Int J Comput Vis
DOI 10.1007/s11263-015-0828-7

Midrange Geometric Interactions for Semantic Segmentation

Constraints for Continuous Multi-label Optimization

Julia Diebold[1] · Claudia Nieuwenhuis[2] · Daniel Cremers[1]

Received: 1 June 2014 / Accepted: 15 May 2015

Abstract In this article we introduce the concept of midrange geometric constraints into semantic segmentation. We call these constraints 'midrange' since they are neither global constraints, which take into account all pixels without any spatial limitation, nor are they local constraints, which only regard single pixels or pairwise relations. Instead, the proposed constraints allow to discourage the occurrence of labels in the vicinity of each other, e.g., 'wolf' and 'sheep'. 'Vicinity' encompasses spatial distance as well as specific spatial directions simultaneously, e.g., 'plates' are found directly above 'tables', but do not fly over them. It is up to the user to specifically define the spatial extent of the constraint between each two labels. Such constraints are not only interesting for scene segmentation, but also for part-based articulated or rigid objects. The reason is that object parts such as for example arms, torso and legs usually obey specific spatial rules, which are among the few things that remain valid for articulated objects over many images and which can be expressed in terms of the proposed midrange constraints, i.e. closeness and/or direction. We show, how midrange geometric constraints are formulated within a continuous multi-label optimization framework, and we give a convex relaxation, which allows us to find globally optimal solutions of the relaxed problem independent of the initialization.

Keywords Variational · Image segmentation · Convex optimization · Directional relations · Geometric relations · Midlevel range interactions

1 Introduction

Semantic segmentation denotes the task of segmenting and recognizing objects based on class-specific information and/or knowledge of typical object relations. Ultimately, we aim at assigning an object label from a given pool of labels to each pixel in the image. In contrast to common segmentation problems, where little or no prior information is available, semantic segmentation makes use of knowledge such as color models, geometric relationships or the likelihood of object constellations, which can be learned from training data. Based on such information we can increase the accuracy of segmentation results and at the same time recognize specific objects instead of only detecting their boundaries.

Especially, the task of segmenting articulated objects is difficult. Animals usually share some common color or texture model, but humans usually wear variable clothes, which makes them hard to segment. Shape priors are often suited well to describe such objects, but they are usually too rigid and do not allow for large pose variations or occlusions. Besides, they are challenging for optimization due to their long-range relations between pixels leading to high-order potentials. We believe that constraints such as geometric relations between objects are generic enough to describe a wide range of objects and poses and still limit the ambiguity of color and texture models, features or object detectors, which usually operate on a single pixel or very limited pixel context.

Communicated by Nikos Komodakis.

✉ Julia Diebold
julia.diebold@tum.de

Claudia Nieuwenhuis
cnieuwe@berkeley.edu

Daniel Cremers
cremers@tum.de

[1] Technische Universität München, Munich, Germany

[2] ICSI, UC Berkeley, Berkeley, USA

Published online: 10 July 2015

Previous optimization approaches for semantic segmentation mainly make use of two types of constraint ranges: local or global ones. Local constraints are usually formulated on a pixel or pairwise pixel level, e.g., color likelihood constraints only consider the deviation of the local pixel color and the precomputed model. In contrast, global constraints are formulated based on the whole image, e.g., size (Möllenhoff et al. 2013; Nieuwenhuis et al. 2013) and volume constraints (Toeppe et al. 2010, 2013) or co-occurrence priors (Ladicky et al. 2010; Souiai et al. 2013b). What has been less explored so far in the context of optimization approaches are midlevel range interactions, i.e., interactions between pixels which are locally confined to a specific user-defined region around each pixel of a specific size, shape and direction.

We see mainly three fields of application of our novel constraints. First, there is the task of scene understanding, where geometric information is very useful to assign correct labels, e.g., knowing that ‘sky’ lies above ‘ground’, that ‘wolf’ and ‘sheep’ usually do not occur together or that ‘boats’ are usually surrounded by ‘water’. Second, there is the task of segmenting objects which consist of several parts, e.g., humans consist of ‘head’, ‘arms’, ‘legs’ and ‘torso’, or cars consist of ‘windshield’, ‘doors’, ‘headlights’, ‘bumpers’ and ‘tires’. For such objects there usually exist specific relations between their parts concerning their location, size and distance. Third, there are scenarios, where we have very specific knowledge of where different objects are located with respect to each other, e.g., when segmenting human clothes. There are no specific object parts, but specific rules about relative positions, and many labels can be missing in contrast to parts of objects.

In all three scenarios, the integration of geometric information into semantic segmentation will improve the labeling results, see Fig. 1 for an example. The main challenge in this article is the formulation and efficient solution of a convex energy optimization problem, which allows for the integration of such additional geometric constraints.

1.1 Related Work

There has been growing interest in the topic of semantic segmentation in recent years, which combines different disciplines such as object detection, various features, shape priors, scene context information and learning. Especially the joint handling of several tasks such as segmentation, recognition and scene classification is beneficial for achieving results of higher quality, but has only recently been made possible by increased computing power.

The typical pipeline of such systems is the following: in the first step, some object detection, region segmentation or superpixel algorithm is used to obtain basic region proposals. In a second step different features are computed from these proposals, which are finally fed into a object classifier such as a random forest, a support vector machine or a neural network (e.g., Carreira and Sminchisescu 2012).

For example, in Arbelaez et al. (2012), combine object detectors, poselets and different features such as color, shape and texture to a powerful semantic segmentation system, which can handle articulated objects in particular. The power of employing millions of features within a random forest approach was demonstrated by Fröhlich et al. (2012). To learn such complex feature hierarchies from large amounts of training data, deep learning was used by Girshick et al. (2014). Instead of non-linear classifiers, Carreira et al. (2012) demonstrated that second order statistics in conjunction with linear classifiers improve semantic segmentation results. A holistic approach to semantic segmentation and the full scene understanding problem which also includes geometric relations such as location or the spatial extent of objects or the type of scene was given by Yao et al. (2012).

In contrast to this typical pipeline processing, we aim at formulating a single optimization problem, which contains all information we have within a single energy. In this way we will be able to guarantee optimality bounds of the solution. To avoid ambiguous solutions which depend on the initialization we will give a convex relaxation of the energy.

The particular novelty of this article in contrast to previous discrete or continuous optimization approaches to semantic segmentation (Bergbauer et al. 2013; Delong and Boykov 2009; Ladicky et al. 2010; Nieuwenhuis et al. 2013; Nosrati et al. 2013; Souiai et al. 2013a, b; Strekalovskiy et al. 2012) is the introduction of midrange geometric constraints between regions concerning relative location, distances and directions.

Ladicky et al. (2010) and Souiai et al. (2013b) introduced co-occurrence priors into semantic segmentation which penalize the simultaneous occurrence of specific label combinations within the same image. In contrast to our approach, these constraints do not consider any spatial information such as location, direction or distance of objects. Ladicky et al. (2010) and Souiai et al. (2013b) model co-occurrence by an additional cost function which can be seen as potentials of the highest order. MRF algorithms for high-order vision problem include Kohli et al. (2007), Komodakis and Paragios (2009). While higher order potentials are generally hard to optimize, the proposed approach is of order two and can be relaxed to a convex optimization problem which can be optimized with standard methods.

Strekalovskiy et al. (2012) took in a way the opposite path and only penalize directly adjacent label combinations. It can be understood as a highly local co-occurrence prior. As the geometric relations in this approach are limited to directly adjacent pixels, they do not include distances or directions. In contrast to previous methods, the method does not require the distance penalty to be a metric but allows combinations which do not adhere to the triangle inequality. While labels

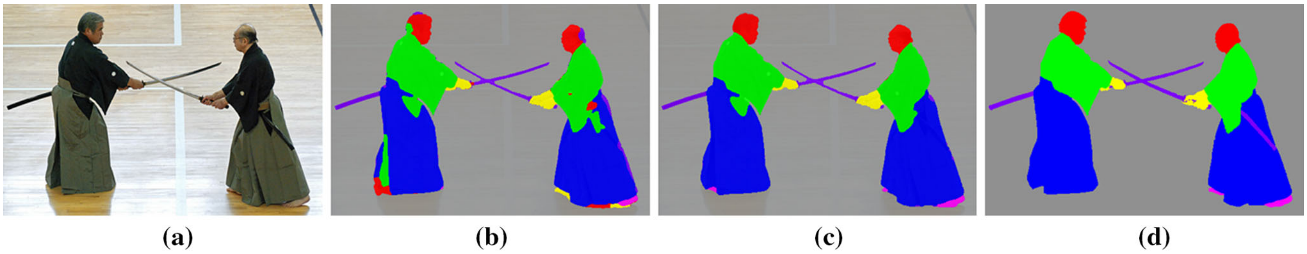


Fig. 1 Midrange geometric constraints improve semantic segmentation results. Midrange geometric constraints between labels allow the user to define specific spatial regions (by means of orientation and distance) within which constraints are enforced, i.e. specific label combinations are penalized. These constraints improve segmentation results, e.g., by imposing penalties for the head being below the jacket or for head and hands being close to the trousers. **a** Original image, **b** Color-based segmentation, **c** Segmentation with novel priors, **d** Ground truth segmentation

'wolf' and 'grass', for example, are common within an image and labels 'sheep' and 'grass' as well, sheep are rarely found next to wolves, which violates the triangle inequality. This often leads to one pixel wide ghost regions of hallucinated objects, which make transitions between two regions cheaper. Our approach does not suffer from ghost regions since our definition of neighborhood regards a larger number of pixels, which makes ghost regions very expensive.

Another type of global constraints for semantic segmentation are hierarchical constraints, which were introduced by Delong et al. (2012) and Souiai et al. (2013a) and penalize the occurrence of objects from semantically different groups or scenes. Constraints relating different region sizes, e.g., of object parts, were introduced by Nieuwenhuis et al. (2013). These constraints are also global and integrate a notion of proportion and size into the segmentation, but they do not take into account distance or directional relations such as that the head of a person usually is above the body.

Topological constraints, which require that some label lies within another label, were proposed within a discrete optimization framework by Delong and Boykov (2009) and within a continuous optimization framework by Nosrati et al. (2013). Geometric scene labeling has been studied by Felzenszwalb and Veksler (2010) considering labelings that have a tiered structure. So-called ordering constraints, which require labels to only occur within a certain direction of other labels, were applied to geometric scene labeling by Liu et al. (2010) for a specific five-part model. Strekalovskiy and Cremers (2011) unified the existing approaches such as the five-regions and the tiered layout and proposed generalized ordering constraints. None of these constraints include any notion of label distance and thus cannot be considered as midrange constraints due to their global nature.

Finally, relative location based geometric relations have been introduced before into segmentation. In Gould et al. (2008) the authors propose a two-stage process, which first uses an appearance model to assign labels and then employs a relative location prior based on the most likely label for each pixel in the first step to improve the segmentation. In contrast to our approach, this is a two-stage process and thus does not allow for any optimality guarantees. In the context of learning, relative spatial label distances have also been successfully applied, e.g., in Kontschieder et al. (2013) and Savarese et al. (2006).

1.2 Contributions and Organization

In this article, we show how midrange geometric constraints characterized by label direction and distance can be integrated into variational semantic segmentation approaches. We give a convex relaxation of the energy minimization problem, which can be solved with fast primal-dual algorithms (Pock et al. 2009) in parallel on graphics hardware (GPUs). Results on various images and benchmarks show that the novel constraints improve semantic segmentation results.

The article is organized as follows: In Sect. 2 we give a formal definition of the multi-label segmentation problem together with different appearance models. In Sect. 3 we introduce the novel midrange geometric priors followed by a convex relaxation of the optimization problem in Sect. 4. In Sect. 5 we present results on various datasets and compare our segmentation results to state-of-the-art approaches.

1.3 Extensions and Improvements over the Previously Published Variant of our Model (Bergbauer et al. 2013)

This journal paper extends our previously published ICCV workshop paper (Bergbauer et al. 2013) by a more general formulation of the proximity priors to midrange geometric constraints and by more detailed and thorough evaluations on various image datasets.

The novel midrange geometric constraints are beneficial for the segmentation of part-based articulated and for part-based rigid objects as well as for the segmentation of scenes. The novel formulation allows to define different structuring elements for each label in contrast to only a single one in

Bergbauer et al. (2013) (Sects. 3.2, 3.3). We give an overview of different appearance models (Sect. 2.2), an analysis of different choices of structuring elements and the penalty matrix (Sects. 3.3, 3.4) as well as a detailed explanation of the impact of different structuring elements (Sect. 3.5).

Additionally, we provide extensive evaluations including failure cases in Sect. 5. We present additional experiments on part-based articulated (Sect. 5.1) and part-based rigid objects (Sect. 5.2) on the CMU-Cornell iCoseg dataset (Batra et al. 2010), the People dataset (Ramanan 2006) and the Penn-Fudan pedestrian database (Wang et al. 2007). Moreover, we show results for the recognition of facades on the eTRIMS image database (Korc and Förstner 2009) and for the task of geometric class labeling of indoor images (Liu et al. 2010) (Sect. 5.3). Furthermore, we provide detailed insights about our experiments including the chosen parameters such as the structuring element $\mathcal{S}_i$ for label i, its size d and the choice of the penalty matrix A.

2 Variational Multi-label Segmentation

We begin with the formal definition of the multi-label segmentation problem and show several choices for the appearance term.

Although any numerical algorithm used for the implementation of the method presented below requires a discretization of the image domain, our general multi-label segmentation framework can be formulated continuously. We present the continuous setup below and give more details regarding the discretization and implementation in Sect. 4.2.

2.1 The Multi-label Optimization Problem

Let $I : \Omega \to \mathbb{R}^d$ denote the input image defined on the image domain $\Omega \subset \mathbb{R}^2$. The general multi-label image segmentation problem with $n \geq 1$ labels consists of the partitioning of the image domain Ω into n regions $\{\Omega_1, \ldots, \Omega_n\}$. This task can be solved by computing binary labeling functions $u_i : \Omega \to \{0, 1\}$ in the space of functions of bounded variation (BV) such that $\Omega_i = \left\{x \,\middle|\, u_i(x) = 1\right\}$. The BV space is important, since it allows jumps in the indicator functions which correspond to sharp transitions between adjacent regions. We compute a segmentation of the image by minimizing the following energy (Zach et al. 2008) (see Nieuwenhuis et al. 2013 for a detailed survey and code)

$$E(\Omega_1, \ldots, \Omega_n) = \frac{\lambda}{2}\sum_{i=1}^{n} \mathrm{Per}_g(\Omega_i) + \sum_{i=1}^{n} \int_{\Omega_i} f_i(x)\, dx. \quad (1)$$

f_i denotes the appearance model for the respective region Ω_i. Different ways to define f_i are discussed in Sect. 2.2. $\mathrm{Per}_g(\Omega_i)$ denotes the perimeter of each set Ω_i, which is minimized in order to favor segments of shorter boundary. These boundaries are measured with either an edge-dependent or a Euclidean metric defined by the non-negative function $g : \Omega \to \mathbb{R}^+$. For example,

$$g(x) = \exp\left(-\frac{|\nabla I(x)|^2}{2\sigma^2}\right), \quad \sigma^2 = \frac{1}{|\Omega|}\int_\Omega |\nabla I(x)|^2 dx$$

favors the coincidence of object and image edges.

To rewrite the perimeter of the regions in terms of the indicator functions we make use of the total variation and its dual formulation (Pock and Chambolle 2011; Nieuwenhuis et al. 2013):

$$\mathrm{Per}_g(\Omega_i) = \int_\Omega g(x)|Du_i| = \sup_{\xi_i : |\xi_i(x)| \leq g(x)} -\int_\Omega u_i \operatorname{div} \xi_i \, dx.$$

Since the binary functions u_i are not differentiable Du_i denotes their distributional derivative. Furthermore, $\xi_i \in \mathcal{C}_c^1\left(\Omega; \mathbb{R}^2\right)$ are the dual variables and $\mathcal{C}_c^1$ denotes the space of smooth functions with compact support. We can rewrite the energy in (1) in terms of the indicator functions $u_i : \Omega \to \{0, 1\}$ (Zach et al. 2008; Nieuwenhuis et al. 2013):

$$\begin{aligned} &E(u_1, \ldots, u_n) \\ &\quad = \sup_{\xi \in \mathcal{K}} \sum_{i=1}^{n} \int_\Omega (f_i - \operatorname{div} \xi_i)\, u_i \, dx, \\ &\quad \text{where } \mathcal{K} = \left\{\xi \in \mathcal{C}_c^1\left(\Omega; \mathbb{R}^{2\times n}\right) \,\middle|\, |\xi_i(x)| \leq \frac{\lambda g(x)}{2}\right\}. \quad (2) \end{aligned}$$

2.2 Choices of Appearance Models

In this article, we use different appearance models for the appearance term f_i in (2) depending on the task to solve.

2.2.1 Color Likelihoods

The simplest model is based on an estimated color probability distribution, e.g., by means of Parzen density estimators. Given a set of scribbles or training data we can extract RGB or HSV sample data for each label in the image or database. A Parzen density for a specific object class i with m_i color samples, each denoted by $I_{ij} \in \mathbb{R}^3$, is then given by

$$\begin{aligned} f_i(x) &:= -\log P_i(I(x)) \\ &:= \frac{1}{m_i}\sum_{j=1}^{m_i} \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} \exp^{-\left((I-I_{ij})^T \Sigma^{-1} (I-I_{ij})\right)}. \end{aligned} \quad (3)$$

The density depends on the covariance matrix Σ of the multivariate Gaussian, which is usually a diagonal matrix and can be adapted by the user. Large values on the diagonal will assign a higher probability to less common colors. Low values on the diagonal will, in contrast, make the distribution more peaked. $|\Sigma|$ denotes the determinant of Σ. In order to avoid infinity values in the appearance term caused by color probabilities of 0 we modify the expression as follows

$$f_i(x) := -\log\Big(P_i\big(I(x)\big)\cdot(1-\epsilon)+\epsilon\Big), \tag{4}$$

where ϵ is a very small constant close to 0.

2.2.2 Spatially Varying Color Likelihoods

In the case of scribble based segmentation we can make use of additional spatial information to estimate color likelihoods. The idea is that close to the scribble we are quite certain about the color in this location, which will be similar to the closest scribble points. On the contrary, far from the scribbles we have to deal with uncertainty in the color density estimation. This level of confidence depends on the distance to the closest scribble point of the current label. It can be integrated into the Parzen density estimator in (3) by computing a different covariance matrix $\Sigma_i(x)$ at each pixel proportional to the distance to the closest scribble of label i:

$$\Sigma_i(x) = \alpha \min_{j=1,\ldots,m_i} |x - x_{ij}|_2, \tag{5}$$

where $\{x_{ij},\ j=1,\ldots,m_i\}$ is the set of user scribbles for region i and $\alpha \in \mathbb{R}$. This yields a space-dependent color density estimator. Details of this approach are given in Nieuwenhuis and Cremers (2013).

2.2.3 Texton Likelihoods

In order to integrate not only color, but also shape and context information, Shotton et al. (2006) proposed to learn a discriminative model to distinguish between object classes. This model is based on texton features, which incorporate shape and texture information jointly. Training is done by means of a shared boosting algorithm. Using the softmax function, the predicted confidence values $H_i(x)$ can be interpreted as a probability distribution. By taking the negative logarithm, we obtain the appearance model

$$f_i(x) = -\log\left(\frac{\exp(H_i(x))}{\sum_{j=1}^{n}\exp\big(H_j(x)\big)}\right), \tag{6}$$

which is also known as unary pixel potential. This model is computed with the ALE library (Ladicky et al. 2009, 2010) and used for the experiments on the Penn-Fudan, eTRIMS and MSRC dataset in order to guarantee comparability to other approaches.

3 The Novel Midrange Geometric Priors

We motivate the midrange geometric priors by means of the simple artificial teddy bear example in Fig. 2a. Common segmentation approaches group pixels mainly according to their color, hence the ears of the bear are associated with the region 'shoes' (Fig. 2b). The desired result, however, would rather connect the ears to the head instead of the shoes as shown in Fig. 2c.

To obtain the desired solution, we make use of a *dilation*, an operation from mathematical morphology. To examine if two regions are close to each other in a certain direction we dilate one of the regions in this direction and compute the overlap between the dilation and the second region. For the teddy example, we want to penalize that head and shoes are close without considering any specific direction. Therefore, we enlarge the region 'shoes' in all directions simultaneously and compute the overlap with the region 'head' as shown in Fig. 2. In this way, we do not only consider directly neighboring pixels as close but we consider proximity with respect to arbitrary neighborhoods of any size, shape or direction, which allows us to introduce midrange geometric constraints. The size and shape of these neighborhoods is determined by the *structuring element* of the dilation and can thus be easily adapted.

3.1 A Continuous Formulation of the Dilation

Dilation is one of the basic operations in mathematical morphology. Since we ultimately aim at introducing the dilation operation into a continuous energy optimization problem instead of using a suboptimal two-step procedure, we require a continuous formulation of the dilation, which can be defined as follows:

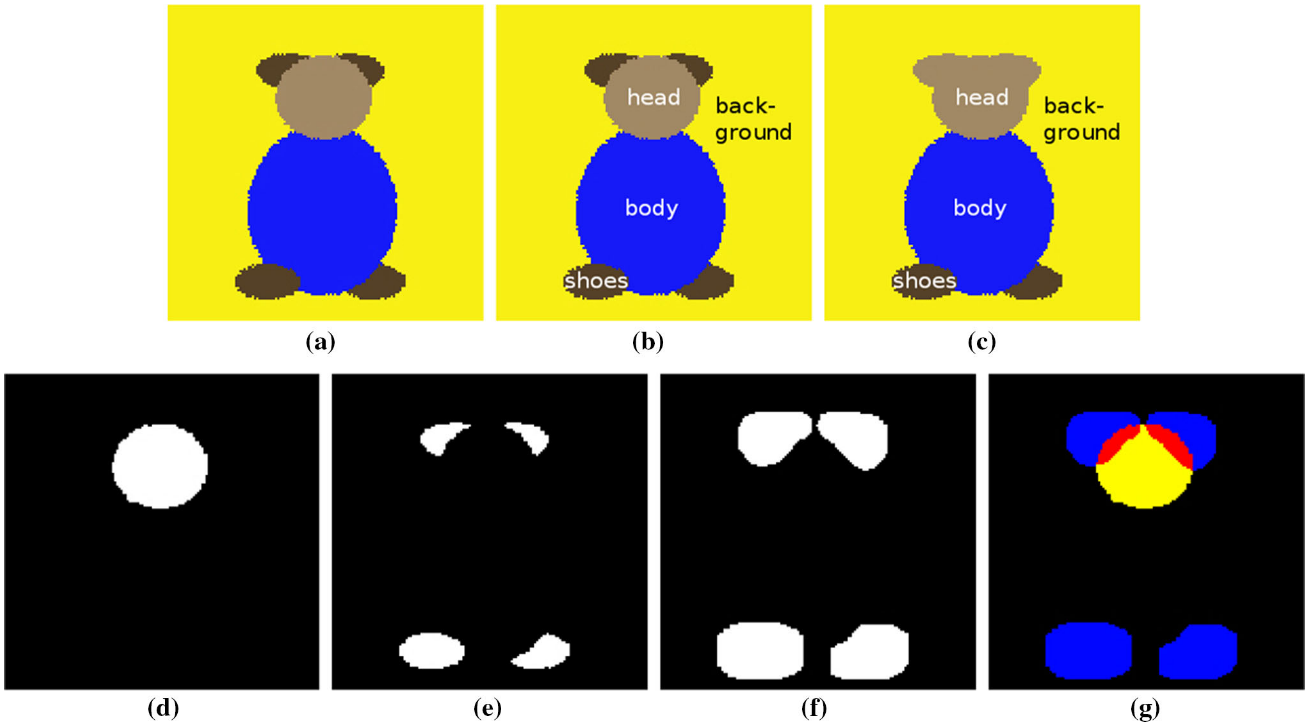
Definition 1 (*Dilation of an image Soille 2003*) Let $I : \Omega \to \mathbb{R}^d$ be an image and $\mathcal{S}$ a structuring element. The dilation of I by $\mathcal{S}$ is denoted by $\delta_{\mathcal{S}}(I)$. The dilated value at a given pixel $x \in \Omega$ is given as follows:

$$[\delta_{\mathcal{S}}(I)](x) = \sup_{z\in\mathcal{S}} I(x+z). \tag{7}$$

Thus, the dilation result at a given location x in the image is the maximum value of the image within the window defined by the structuring element $\mathcal{S}$, when its origin is at x.

3.2 Introducing Midrange Geometric Constraints

To compute the proximity of two labels, we first introduce the notion of an extended region indicator function u_i denoted by $d_i : \Omega \to \{0,1\}$, which dilates the indicator function in a specific direction and distance (see Fig. 2 and Definition 1):



(a) (b) (c)

(d) (e) (f) (g)

Fig. 2 Introducing midrange geometric priors. *First row* **a** Original, **b** Color-based segmentation, **c** Desired segmentation. Color-based segmentation often fails. The ears of the bear are (**b**) assigned the label 'shoes' instead of (**c**) being combined with the label 'head'. *Second row* The novel priors can be used to penalize the 'closeness' of two labels, in this example **d** 'head' and **e** 'shoes'. **f** Dilation of the indicator function 'shoes'; **g** Overlap of the dilated region 'shoes' (*blue*) and the region 'head' (*yellow*). Appropriate penalties for such overlap (*red*) introduce semantic information into the segmentation

$$d_i(x) := \left[\delta_{\mathcal{S}_i}(u_i)\right](x) = \sup_{z\in\mathcal{S}_i} u_i(x+z). \tag{8}$$

The set $\mathcal{S}_i$ determines the type of geometric spatial relationship we want to penalize for label i, i.e., distance and direction, for example 'less than 20 pixels above'. $\mathcal{S}_i$ is often denoted by *structuring element*. We will give a more detailed explanation of $\mathcal{S}_i$ in the Sect. 3.3.

To detect if two regions i and j are close to each other, we compute the overlap of the extended indicator function d_i and the indicator function u_j, as shown in bright red in Fig. 2g. For each two regions i and j we can now penalize their proximity by means of the following energy term:

$$E_{geom}(u) = \sum_{1\le i<j\le n} A(i,j)\int_{\Omega} d_i(x)\,u_j(x)\,dx. \tag{9}$$

The penalty matrix $A \in \mathbb{R}^{n\times n}_{\geq 0}$ indicates the penalty for the occurrence of label j in the proximity of label i. Information on how to define or learn this matrix are given in Sect. 3.4.

3.3 Structuring Elements

The dilation operation requires a *structuring element* (SE) for probing and expanding label indicator functions. The option to use structuring elements of different sizes and shapes is one of the major benefits of the proposed algorithm.

There are many different ways to define SEs. We can specify one set $\mathcal{S}_i$ for each label i. If $\mathcal{S}_i$ is for example a line we can penalize the proximity of specific labels in specific directions, e.g., the occurrence of a book below a sign (compare Fig. 4c). Symmetric sets of specific sizes consider the proximity of two labels without preference of a specific direction. Sparse sets $\mathcal{S}_i$ as shown in Fig. 3c and d lead to similar results but can speed up the runtime. Examples for structuring elements are shown in Fig. 3 and their application in Fig. 4.

The larger $\mathcal{S}_i$ the more pixels are considered adjacent to x. Let the occurrence of label j in the proximity of label i be denoted by $i \sim_{\mathcal{S}_i} j$. If training data is available we can learn the probabilities $P\left(i \sim_{\mathcal{S}_i} j\right)$ for different types and sizes of SEs and then define $\mathcal{S}_i$ as the SE which provides the highest information gain for label i.

The information gain for a label i and structuring element $\mathcal{S}$ can be computed by means of the Shannon entropy (Shannon 2001):

$$H(i,\mathcal{S}) = -\sum_{\substack{j\in\{1,\dots,n\}\\ j\neq i}} P\left(i \sim_{\mathcal{S}} j\right)\cdot\log\left(P\left(i \sim_{\mathcal{S}} j\right)\right). \tag{10}$$

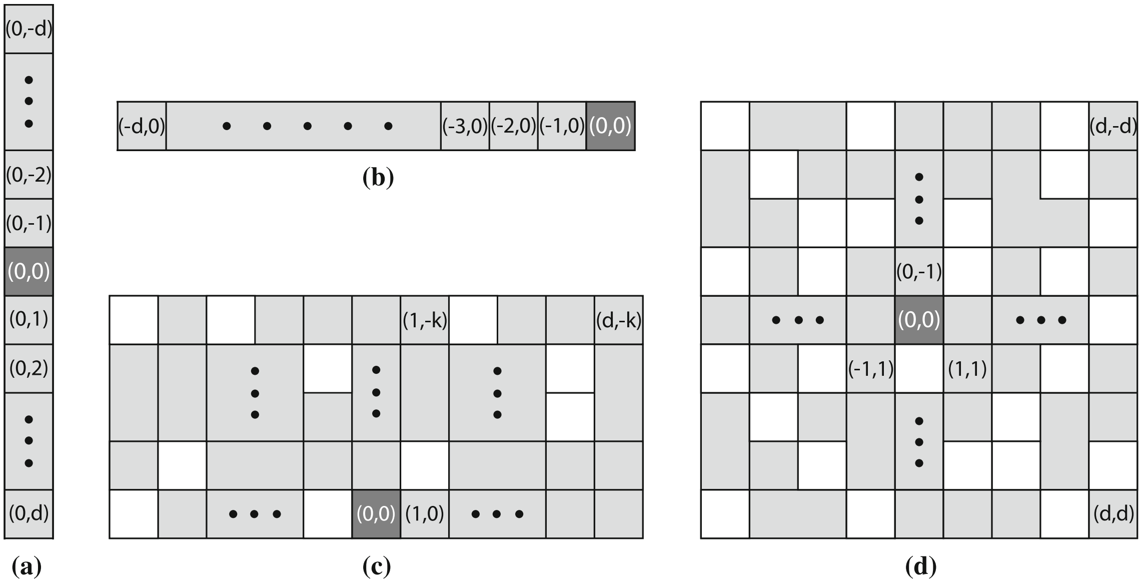


Fig. 3 Horizontal, vertical and sparse structuring elements. Knowledge of the occurrence of regions *above/below* or *left/right* within a distance *d* can be included by using different structuring elements. Each structuring element has an origin which is indicated in *dark gray*. **a** The vertical line dilates a region *d* pixels upward and downward, **b** the horizontal line centered on the rightmost pixel enlarges a region *d* pixels to the right. **c**, **d** To save computation time sparse structuring elements can be used. White pixels are chosen randomly and left out, i.e. they are not included in the set $\mathcal{S}$. **c** A sparse element, which dilates to the bottom, right and left. **d** A sparse element, which dilates equally in all directions and thus only regards pixel distance

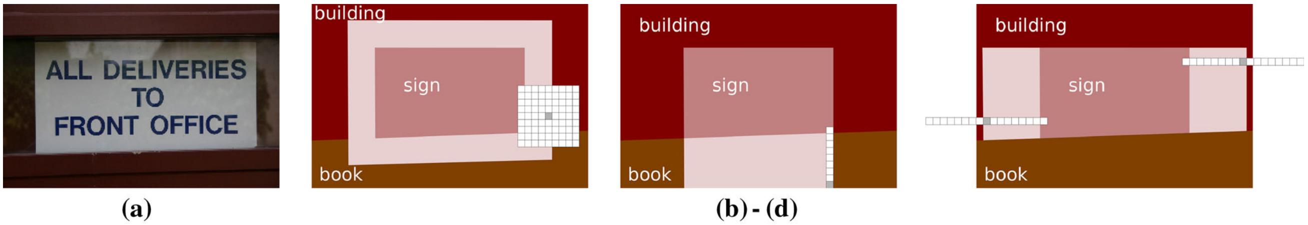


Fig. 4 Impact of structuring elements. **a** Original image, **b–d** Indicator function extended by different sets $\mathcal{S}_i$. The light pink color illustrates the extended 'sign' region. Different sets $\mathcal{S}_i$ convey different geometric priors. **b** Symmetric sets $\mathcal{S}_i$ only consider object distances, but are indifferent to directional relations. **c** If $\mathcal{S}_i$ is chosen as a vertical line centered at the bottom, the indicator function of the region 'sign' is extended to the bottom of the object, e.g., penalizing 'book' appearing closely (within *d* pixels) below 'sign'. **d** *Horizontal lines* penalize labels to the *left* and *right*

The probabilities $P\,(i \sim_{\mathcal{S}} j)$ can e.g., be obtained by estimating the relative frequencies of the labels within the range of the selected structuring element $\mathcal{S}$ in the training data. We can either treat the relative frequencies as a joint probability distribution, which requires normalization by the sum of all elements, or we can treat it as a conditional distribution, which requires normalization per label separately. In the first case, the occurrence probability of each label is inherently part of the estimated probability distribution, i.e., labels occurring rarely in the training data also occur rarely close to other labels. The second case removes the influence of the frequency of label occurrences and only judges if a second label is common within the vicinity of a first label, which is already given. A slightly different way, which does not involve probability distributions, is to count all pairwise label co-occurrences in the training data weighted by their inverse distance in a matrix $B_{\mathcal{S}}$, to normalize $B_{\mathcal{S}}$ and then to estimate $P\,(i \sim_{\mathcal{S}} j)$ by $B_{\mathcal{S}}(i, j)$. For the Penn-Fudan dataset and different types and sizes of SEs for each label (except the 'background'), for example, we use the latter approach and obtain the SEs in Fig. 5.

Note that the optimal structuring element $\mathcal{S}_i$ for label i will be dependent on the viewpoint. According to whether a scene is captured from a front or a top view, the size, shape and position of the objects in the scene varies in the captured image. Hence, to define one structuring element $\mathcal{S}_i$ for all labels i in a benchmark, some uniformity of the training and test images has to be assumed.

If a learning approach is not desired or not possible due to lack of training images or non-uniformity of the dataset, appropriate sets $\mathcal{S}_i$ can easily be chosen manually as done for the experiments in Figs. 9 and 10 in Fig. 7.

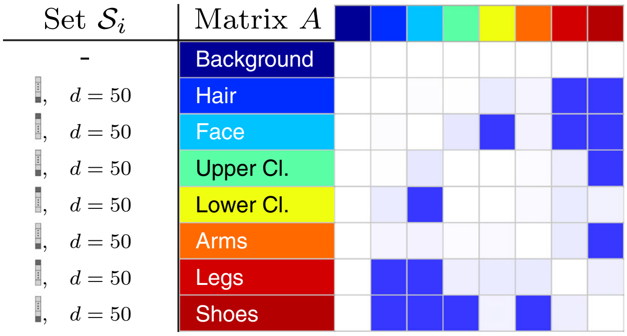


Fig. 5 SEs and penalty matrix A learned on the Penn–Fudan training set (label colors according to benchmark conventions). We penalize the labels 'lower clothes', 'legs' and 'shoes' above 'face', as well as 'hair', 'face', 'upper clothes' and 'arms' below 'shoes'

3.4 Specification of the Penalty Matrix

To introduce the novel geometric priors into the original optimization problem in (2), we have to define the penalty matrix $A \in \mathbb{R}^{n\times n}_{\geq 0}$ in (9). Each entry $A(i, j)$, $i \neq j$ indicates the penalty for the occurrence of label j in the proximity of label i, where the proximity is defined by the respective structuring element $\mathcal{S}_i$. For $i = j$ we set $A(i, i) := 0$.

If training data is available we can learn the probabilities $P\left(i \sim_{\mathcal{S}_i} j\right)$ as described in Sect. 3.3 and define the entries $A(i, j)$ for label j being close to label i, e.g., by $A(i, j) := \min(-\log(P\left(i \sim_{\mathcal{S}_i} j\right)), m)$ with a fixed number $m \in \mathbb{N}$. This assigns a penalty close to zero to frequent and a penalty of m to less frequent co-occurrences. For the MSRC benchmark and a symmetric set $\mathcal{S}$ of size 9 × 9 for all labels, for example, we estimate $P\,(i \sim_{\mathcal{S}} j)$ by $B_{\mathcal{S}}(i, j)$ (compare Sect. 3.3) and define $A(i, j) := \min(-\log(B_{\mathcal{S}}\,(i, j)), 20)$ and obtain the penalty matrix in Fig. 6. The first column in Fig. 6, e.g., indicates that the occurrence of 'building' close to 'tree' or 'sky' is very likely (light colored cells), whereas the occurrence of 'building' close to 'sheep' is very unlikely (dark colored cell).

If there is no appropriate training data available or if a learning approach is not desired, the penalty matrix A can easily be defined by hand as done for the experiments in Figs. 9 and 10 in Fig. 7.

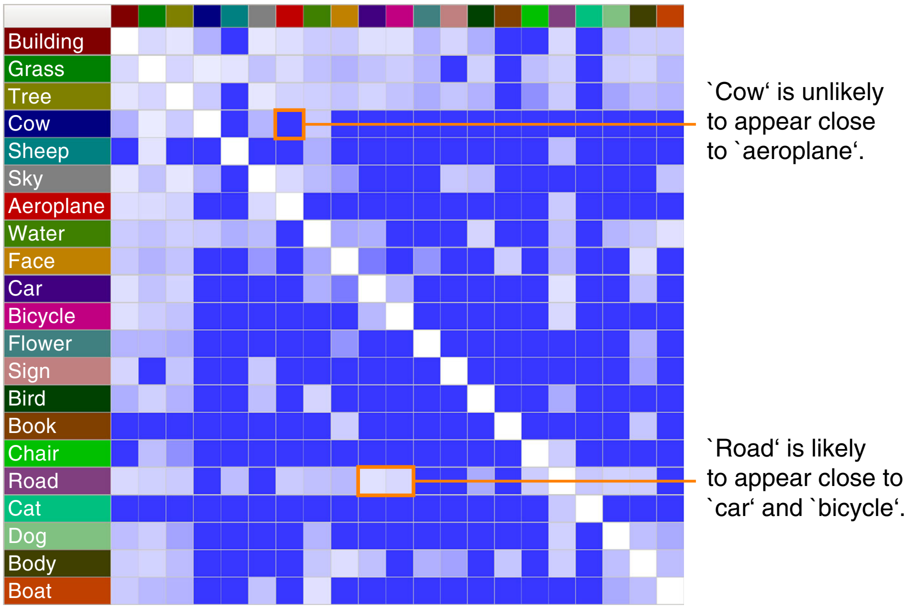
3.5 Real-World Examples

We demonstrate the impact of the novel midrange geometric priors by means of two examples shown in Figs. 8, 9 and 10. The corresponding color-legend can be found in Fig. 11.

Figure 7 gives an overview of the structuring elements and the penalty matrices defined for the segmentation of the soccer player and the fighters. For each label i an individual structuring element $\mathcal{S}_i$ with specific size d has been defined by the user. In the example of the soccer player we penalize the label 'head' being close to 'arms', below 'shirt' or below 'legs', as well as 'arms' below 'shirt' or 'legs'. For the fighters, we penalize the occurrence of 'head' below 'jacket', close to 'trousers' or close to 'weapon'. Furthermore, we penalize 'hands' next to 'trousers'.

Figure 8 shows the generation of the extended indicator functions d_i by means of different structuring elements from the original indicator functions u_i.

Figures 9 and 10 show how segmentation results can be improved by imposing midrange geometric constraints by penalizing the overlap of the specified indicator functions.



`Cow‘ is unlikely to appear close to `aeroplane‘.

`Road‘ is likely to appear close to `car‘ and `bicycle‘.

Fig. 6 Penalty matrix A learned on the MSRC training data (objects are *color coded* corresponding to benchmark convention in *first row* and *column*). The *lighter the color* the more likely is the co-occurrence of the corresponding labels within the relative spatial context, and the lower is the corresponding penalty

Fig. 7 Penalty matrix A and corresponding structuring elements (SE) defined to improve the segmentation results in Figs. 9 and 10. For each label a specific SE with specific size d has been chosen by the user. For each label pair the corresponding matrix entry indicates the penalty in case these labels appear close to each other in the specified direction

Set $\mathcal{S}_i$	Matrix A		🟥	🟨	🟩	🟪	⬜
▦, $d = 15$	Head	🟥	0	12	0	12	0
▦, $d = 20$	Arms	🟨	12	0	0	12	0
▯, $d = 50$	Shirt	🟩	12	12	0	0	0
▤, $d = 20$	Legs	🟪	24	24	0	0	0
-	Background	⬜	0	0	0	0	0

Set $\mathcal{S}_i$	Matrix A		🟥	🟦	🟦	🟨	🟪	⬜	🟪
-	Head	🟥	0	0	0	0	0	0	0
▯, $d = 20$	Jacket	🟦	10	0	0	0	0	0	0
▦, $d = 20$	Trousers	🟦	10	0	0	50	0	0	0
-	Hands	🟨	0	0	0	0	0	0	0
-	Feet	🟪	0	0	0	0	0	0	0
-	Background	⬜	0	0	0	0	0	0	0
▦, $d = 25$	Weapon	🟪	10	0	0	0	0	0	0

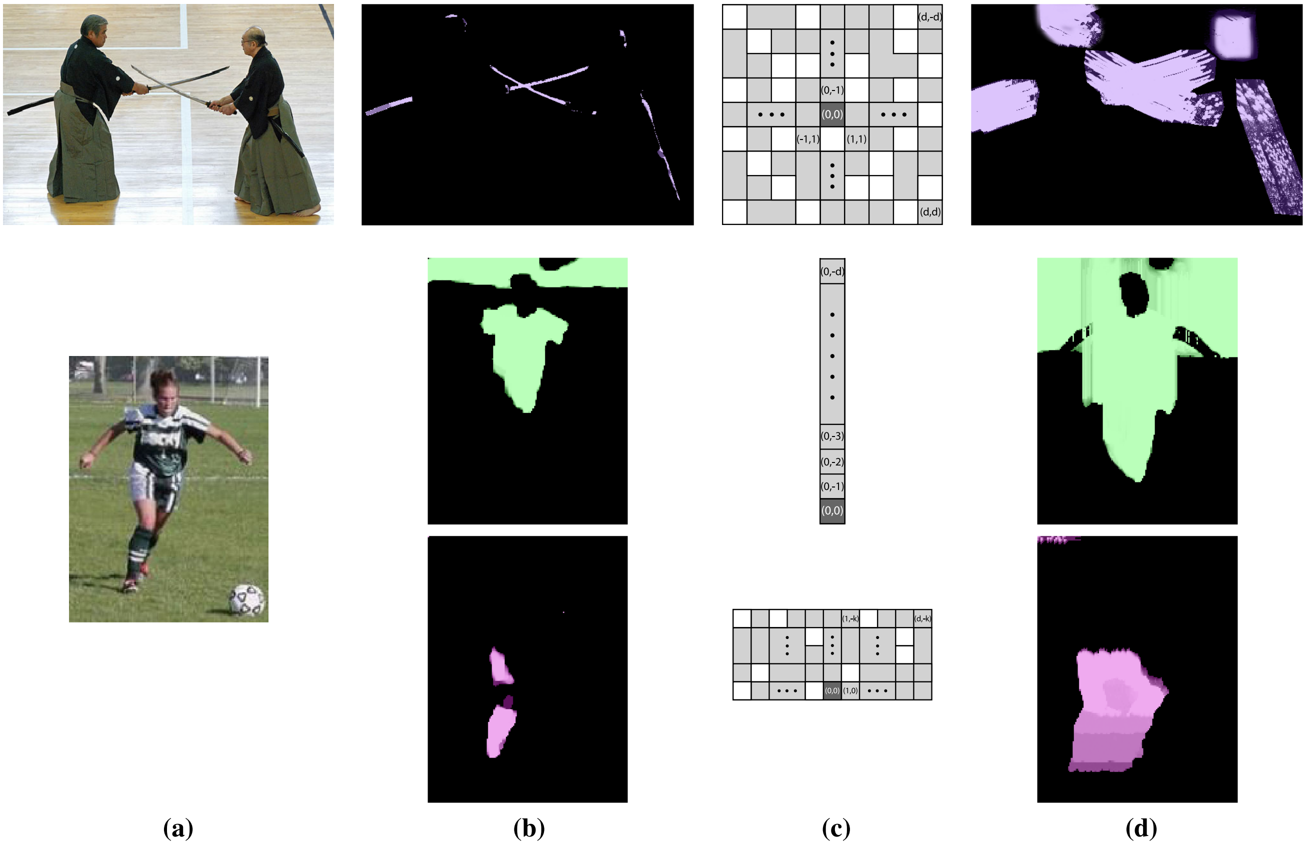


Fig. 8 Effect of different sets $\mathcal{S}_i$ shown by means of the extended region indicator functions d_i. **a** Original images and **b** indicator functions u_i for the 'weapon' region of the fighters and the 'shirt' and 'legs' region of the soccer player. **c** Sets $\mathcal{S}_i$ chosen for the dilation. *Top* Symmetric sets $\mathcal{S}_i$ consider proximity in all directions. *Center* If $\mathcal{S}$ is chosen as a vertical line centered at the bottom, the indicator function of the region 'shirt' is extended to the bottom of the object, e.g., penalizing 'head' appearing below 'shirt'. *Bottom* Horizontal lines penalize labels to the left and right and can be extended to probe *downwards* to the left and right. Sparse sets save runtime. **d** Extended indicator functions $d_i = \delta_{\mathcal{S}_i}(u_i)$ obtained with $\mathcal{S}_i$

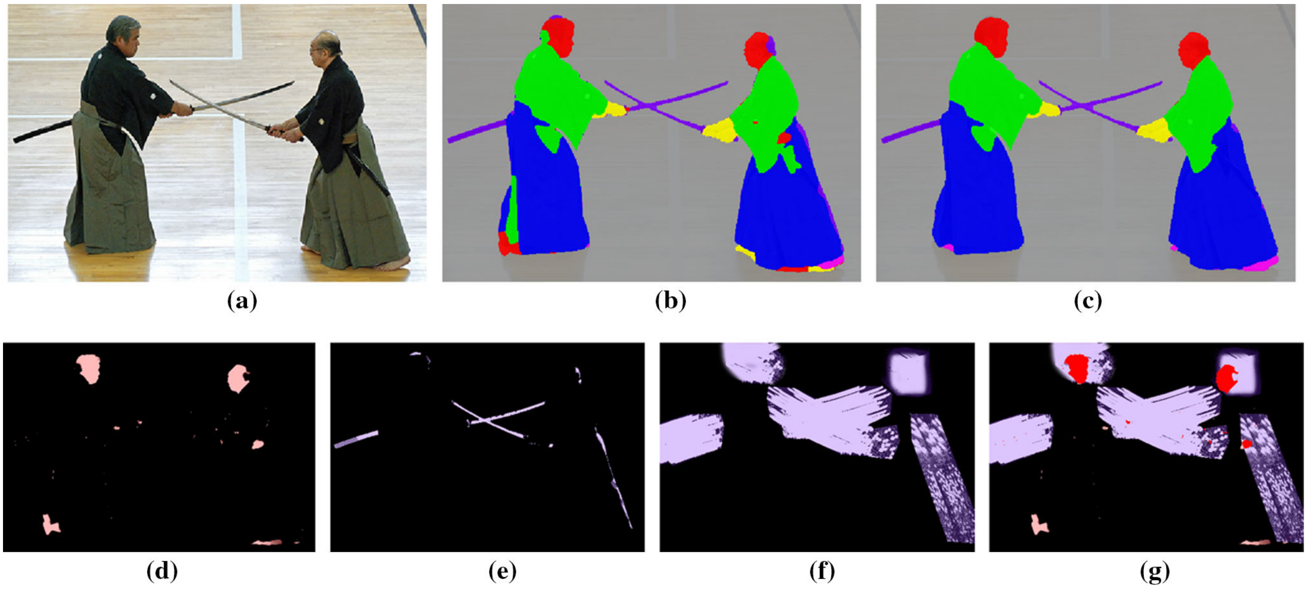


Fig. 9 Penalization of the proximity of the labels ‘head’ and ‘weapon’ for the fighter image **(a)** to improve the color-based segmentation in **(b)**, see Fig. 11 for a color legend. **d, e** Region indicator functions u_i for ‘head’ and ‘weapon’. **f** Extended region indicator function d_i for ‘weapon’. **g** *Bright red* indicates the penalized overlap. The overlap of the head with the weapon forces the weapon to retract in the area of the head. The novel geometric priors use several geometric constraints together and yield the result in **(c)**

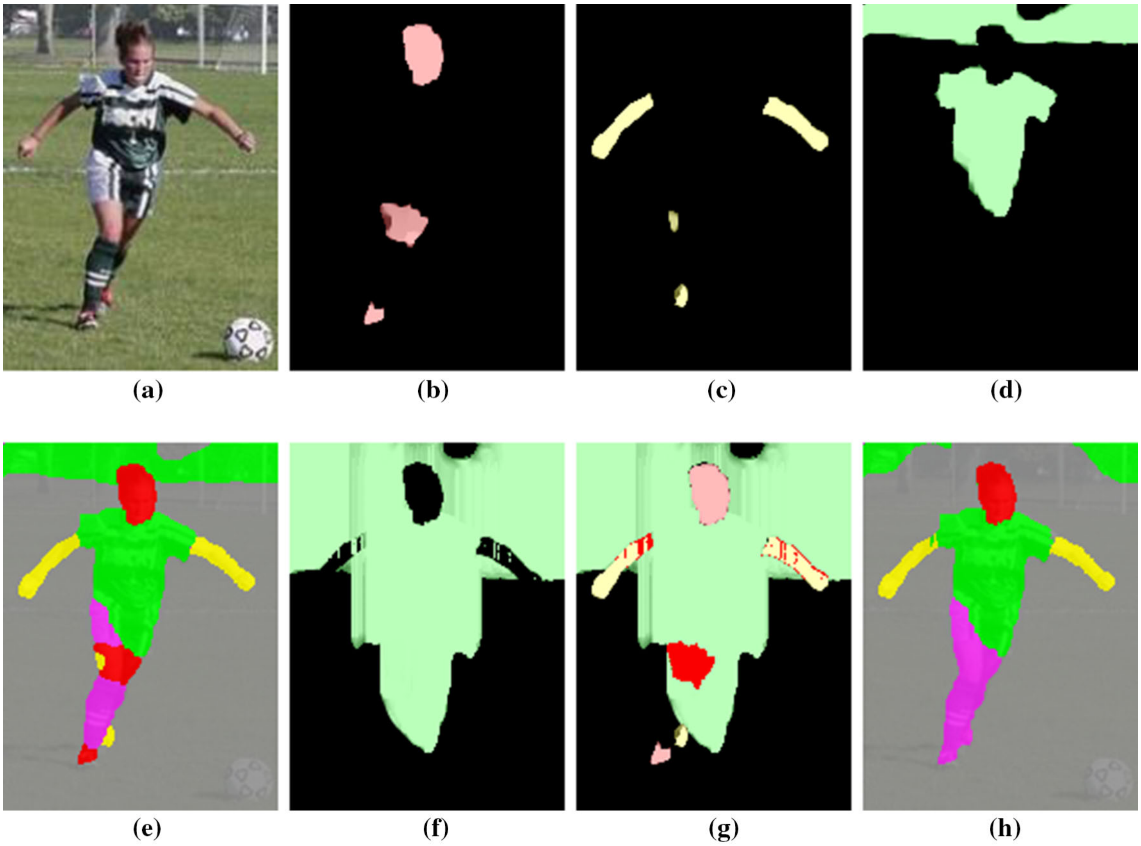


Fig. 10 Effect of novel geometric constraints **(h)**, which improve the color-based segmentation in **(e)** for the soccer player image **(a)**, see Fig. 11 for a color legend. **b–d** Region indicator functions u_i for ‘head’, ‘arms’ and ‘shirt’. **f** Extended indicator function d_i for the shirt region. **g** The overlap of the head and the arms with the dilated shirt force the shirt in the top of the image to retract and the head to disappear from the trousers

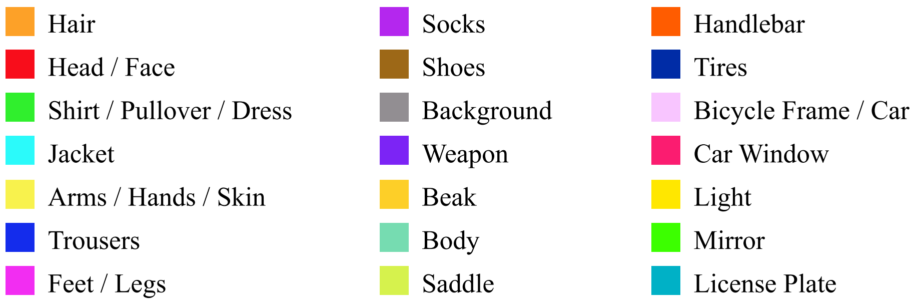


Fig. 11 *Color legend* used in all experiments except for the benchmarks which have their own color coding

4 Integrating the Geometric Constraints into a Convex Optimization Problem

After introducing and defining the novel midrange geometric constraints (9) with $A \in \mathbb{R}^{n\times n}_{\geq 0}$ it remains to integrate these constraints into the original convex optimization problem for segmentation (2)

$$\min_{u\in\mathcal{G}} E(u) + E_{geom}(u) \tag{11}$$

$$= \min_{u\in\mathcal{G}} \sup_{\xi\in\mathcal{K}} \sum_{i=1}^{n} \int_{\Omega} (f_i - \operatorname{div} \xi_i)\, u_i \, dx + \sum_{1\leq i<j\leq n} A(i,j) \int_{\Omega} d_i u_j \, dx \tag{12}$$

$$\text{s.t. } d_i(x) = \left[\delta_{\mathcal{S}_i}(u_i)\right](x) = \sup_{z\in\mathcal{S}_i} u_i(x+z), \tag{13}$$

$$\mathcal{G} = \left\{ u \in BV\left(\Omega; \{0,1\}^n\right) \,\middle|\, \sum_{j=1}^{n} u_j(x) = 1 \ \ \forall\, x \in \Omega \right\}, \tag{14}$$

$$\mathcal{K} = \left\{ \xi \in C_c^1\left(\Omega; \mathbb{R}^{2\times n}\right) \,\middle|\, |\xi_i(x)| \leq \frac{\lambda g(x)}{2} \right\}. \tag{15}$$

4.1 A Convex Relaxation of the Midrange Geometric Constraints

In the following we will propose a convex relaxation of the segmentation problem (2) combined with the proposed priors in (9) as stated in (11)–(15). To obtain a convex optimization problem, we require convex functions over convex domains.

4.1.1 Relaxation of the Binary Functions u_i

The general multi-labeling problem is not convex due to the binary region indicator functions $u_i : \Omega \to \{0,1\}$ in (14). To obtain a convex problem where each pixel is assigned to exactly one label, instead of optimizing over the set $\mathcal{G}$ in (14) optimization is carried out over the convex set

$$\mathcal{U} = \left\{ u \in BV\left(\Omega; [0,1]^n\right) \,\middle|\, \sum_{j=1}^{n} u_j(x) = 1 \ \ \forall\, x \in \Omega \right\}.$$

4.1.2 Relaxation of the Dilation Constraints

The dilation constraints in (13) are relaxed to

$$d_i(x) \geq u_i(x+z) \quad \forall\, x \in \Omega,\ z \in \mathcal{S}_i. \tag{16}$$

By simultaneously minimizing over the functions d_i we can assure that at the optimum d_i fulfills the constraints in (13) exactly. The inequality (16) can easily be included in the segmentation energy by introducing a set of Lagrange multipliers β_{i_z} and adding the following energy term:

$$\min_{d\in\mathcal{D}} \max_{\beta\in\mathcal{B}} \sum_{i=1}^{n} \sum_{z\in\mathcal{S}_i} \int_{\Omega} \beta_{i_z}(x) \left(d_i(x) - u_i(x+z)\right) dx,$$

$$\mathcal{B} = \left\{ \beta_{i_z} \,\middle|\, \beta_{i_z} : \Omega \to [-\infty, 0] \ \ \forall\, z \in \mathcal{S}_i, \quad i = 1, \ldots, n \right\},$$
$$\mathcal{D} = BV\left(\Omega; [0,1]^n\right). \tag{17}$$

4.1.3 Relaxation of the Product of the Indicator Functions

The product of the dilation d_i and the indicator function u_j in (12) is not convex. A convex, tight relaxation of such energy terms was given by Strekalovskiy et al. (2011). To this end, we introduce additional dual variables q_{ij} and Lagrange multipliers α_{ij}:

$$\mathcal{Q} = \left\{ q_{ij} \,\middle|\, q_{ij} : \Omega \to \mathbb{R}^4,\ 1 \leq i < j \leq n \right\},$$
$$\mathcal{A} = \left\{ \alpha_{ij} \,\middle|\, \alpha_{ij} : \Omega \to [-\infty, 0]^4,\ 1 \leq i < j \leq n \right\}. \tag{18}$$

4.1.4 Resulting Optimization Problem

After carrying out these relaxations we finally obtain the following convex energy minimization problem

$$\min_{\substack{u\in\mathcal{U}\\ d\in\mathcal{D}\\ \alpha\in\mathcal{A}}} \max_{\substack{\xi\in\mathcal{K}\\ \beta\in\mathcal{B}\\ q\in\mathcal{Q}}} \sum_{i=1}^{n} \left\{ \int_{\Omega} \left(f_i(x) - \operatorname{div} \xi_i(x)\right) u_i(x)\, dx + \sum_{z\in\mathcal{S}_i} \int_{\Omega} \beta_{i_z}(x) \left(d_i(x) - u_i(x+z)\right) dx \right.$$

$$\begin{aligned}
&+\sum_{j=i+1}^{n}\int_{\Omega} q^1_{ij}(x)\left(1-d_i(x)\right)+q^2_{ij}(x)\,d_i(x)\\
&+q^3_{ij}(x)\left(1-u_j(x)\right)+q^4_{ij}(x)\,u_j(x)\\
&+\alpha^1_{ij}(x)\left(q^1_{ij}(x)+q^3_{ij}(x)\right)+\alpha^2_{ij}(x)\left(q^1_{ij}(x)+q^4_{ij}(x)\right)\\
&+\alpha^3_{ij}(x)\left(q^2_{ij}(x)+q^3_{ij}(x)\right)+\alpha^4_{ij}(x)\left(q^2_{ij}(x)\right.\\
&\left.+q^4_{ij}(x)-A(i,j)\right)dx\Big\}. \qquad (19)
\end{aligned}$$

The projections onto the respective convex sets of ξ, d, β and α are done by simple clipping while that of the primal variable u is a projection onto the simplex in $\mathbb{R}^n$ (Michelot 1986).

4.2 Implementation

In the previous sections, we proposed our method in a continuous framework with the image domain $\Omega \subset \mathbb{R}^2$. For this reason we discretize the problem using a regular Cartesian grid (Chambolle and Pock 2011) as is commonly done, e.g., Pock et al. (2009). In order to find the globally optimal solution to this relaxed convex optimization problem, we employ the primal-dual algorithm published in Pock et al. (2009). Optimization is done by alternating a gradient descent with respect to the functions u, d and α and a gradient ascent for the dual variables ξ, β and q interlaced with an over-relaxation step in the primal variables. The step sizes are chosen optimally according to Pock and Chambolle (2011).

Due to the inherent parallel structure of the optimization algorithm (Pock et al. 2009), each pixel can be updated independently. E.g., the update of the indicator function $u(x)$: $u^n \to u^{n+1}$ can be computed in parallel for each pixel $x \in \Omega$. Hence, the approach can be easily parallelized and implemented on graphics hardware. We used a parallel CUDA implementation on an NVIDIA GTX 680 GPU.

We stopped the iterations when the average update of the indicator function $u(x)$ per pixel was less than 10^{-5}, i.e., if

$$\frac{1}{|\Omega|}\left|u^k - u^{k-1}\right| < 10^{-5}. \qquad (20)$$

By relaxing the indicator variables, i.e., allowing the primal variables u_i to take on intermediate values between 0 and 1, we may end up with non-binary solutions. In order to obtain a binary solution to the original optimization problem, we assign each pixel x to the label L with maximum value after optimizing the relaxed problem:

$$L(x) = \arg\max_i \{u_i(x)\}, \quad x \in \Omega. \qquad (21)$$

We observed that the computed relaxed solutions u are binary almost everywhere. For the benchmark experiments, the computed solutions $u_i(x) < 0.01$ or $u_i(x) > 0.99$ for 97–98 % of all pixels $x \in \Omega$ and $i = 1, \ldots, n$ and for 2–3 % $u_i(x) \in [0.01, 0.99]$.

5 Experiments and Results

We have shown how to integrate midrange geometric priors into a variational multi-label approach and gave a convex relaxation of the resulting optimization problem. One of the major advantages of the proposed algorithm is that we can utilize sets $\mathcal{S}_i$ of different sizes and shapes which allow us to define specific neighborhoods of different spatial extent and direction for each label. In the following we will show qualitative and quantitative results for a number of articulated part-based objects such as humans, animals or clothes from the CMU-Cornell iCoseg (Batra et al. 2010), People (Ramanan 2006) and Penn-Fudan dataset (Wang et al. 2007), for rigid part-based objects such as cars or bicycles as well as for a variety of scenes in the MSRC benchmark, for the recognition of facades on the eTRIMS image database (Korc and Förstner 2009) and for the task of geometric class labeling of indoor images (Liu et al. 2010).

For the iCoseg and People dataset, we defined the labels: ‘hair’, ‘face’, ‘shirt’, ‘jacket’, ‘hands’, ‘trousers’, ‘feet’, ‘socks’, ‘shoes’, ‘weapon’ and ‘background’. The corresponding colors are indicated in Fig. 11 and consistently used for all experiments except for the benchmarks which have their own standard color legends.[1]

5.1 Part-based Articulated Objects: Humans, Animals, Clothes

Articulated objects such as humans, animals and clothes are hard to segment correctly since there are few things that remain constant over a set of images and thus suitable for formulating useful constraints, for example color, shape or absolute location priors are not suitable. Yet, what is typical for many of these objects is that they obey relative geometric constraints, which relate to specific directions and distances and which can be formulated within the proposed framework of the midrange geometric constraints. Especially humans, animals and clothes are good examples for objects, which are difficult to segment, but still follow strict rules imposed on their parts, e.g., the head is usually above the feet and

[1] The Pascal VOC dataset is not appropriate for the evaluation of the proposed midrange geometric priors since the images of the Pascal VOC segmentation task consist of only very few (often only one) objects and large ‘background’ areas. 64 %/90 % of the images contain less or equal one/two objects. The proposed constraints, however, allow to discourage the occurrence of labels in the vicinity of each other, e.g., that ‘sky’ lies above ‘ground’ or that the ‘shoes’ of a person appear below the ‘head’. We therefore chose datasets with more than three labels for the benchmark evaluations.

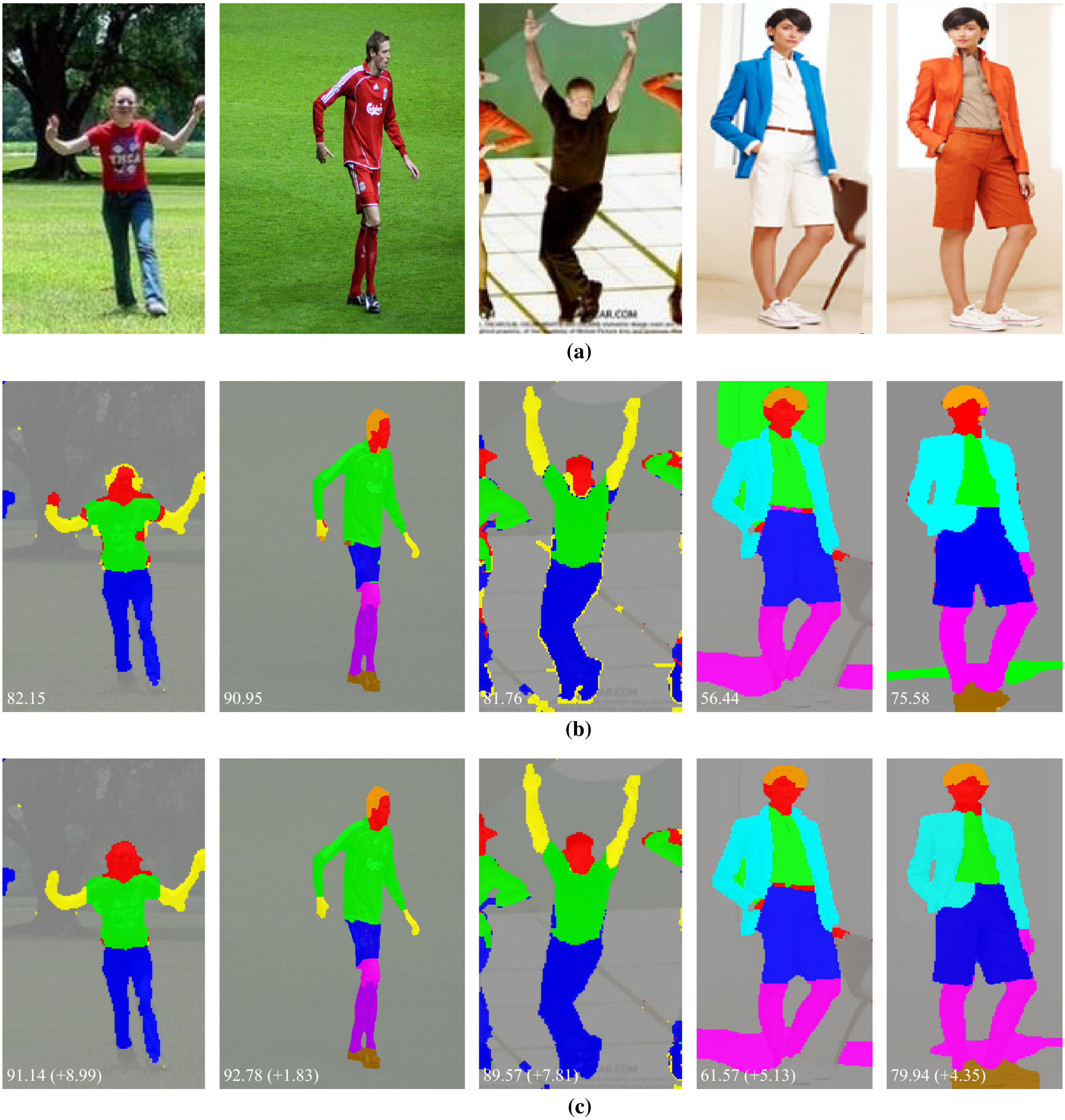
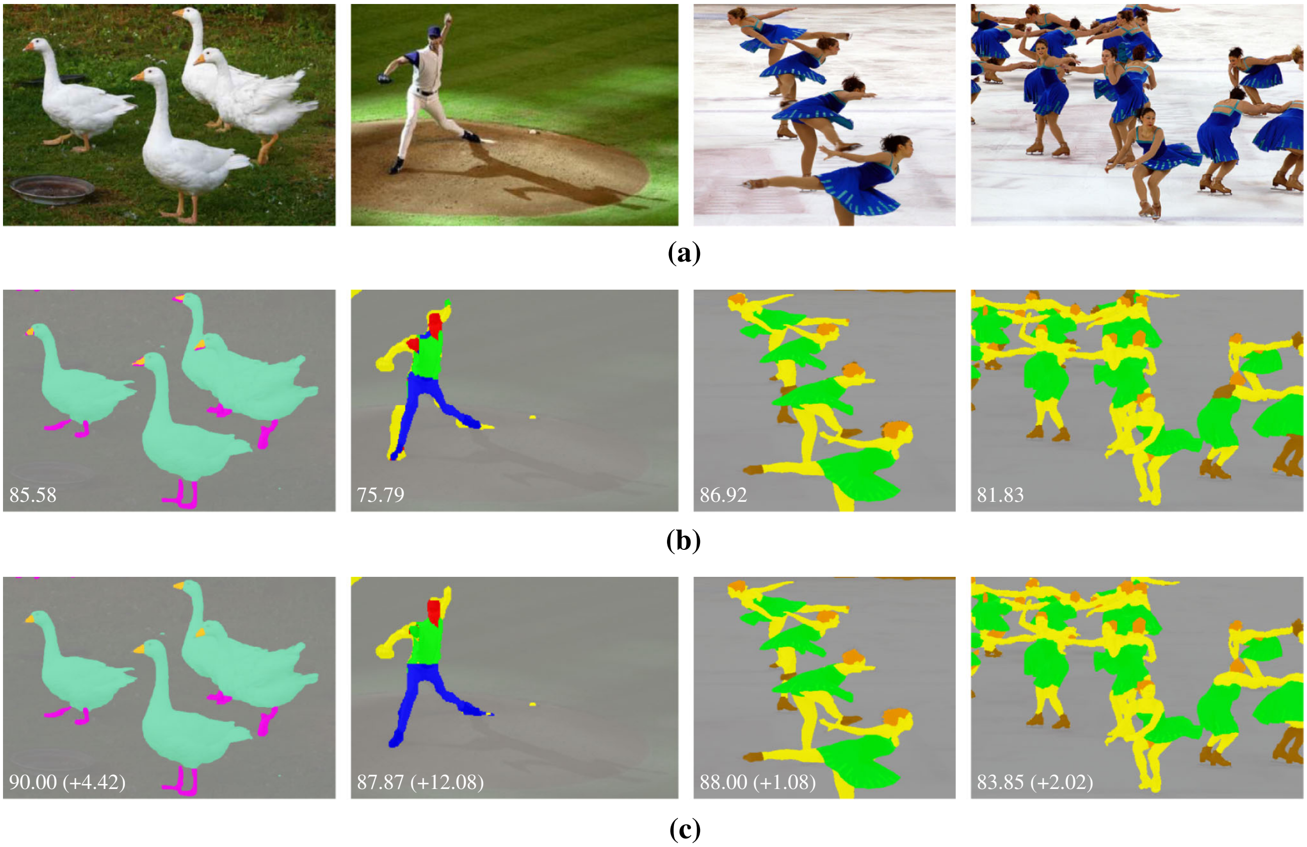


Fig. 12 Part-based articulated objects such as humans or clothes. **a** Original images, **b** Color-based segmentation (solution of Eq. (2)). Improved segmentation results **(c)** can be obtained by introducing the proposed novel midrange geometric constraints in order to introduce prior knowledge of relative location, direction and distance of body parts, e.g., we penalize 'trousers' above 'body', 'head' 'arms' below 'legs' and 'shirt' next to 'shoes'. The dice-score (and the improvement over the color-based segmentation) is given in white in the bottom left image corner

trousers can be found below the shirt and hands are usually close to arms.

Figures 12, 13 and 14 show segmentation results for humans, clothes and animals. Since no training data is available for the iCoseg and People dataset, we manually defined the structuring elements $\mathcal{S}_i$ and the penalty matrix A. For example, we penalize 'arms' and 'trousers' next to one another using a 31×31 sparse symmetric structuring element as well as 'hair' and 'face' next to 'hands' by a 51×51 sparse symmetric element $\mathcal{S}_i$ (compare Fig. 3d for $d = 15, 25$). Furthermore, we penalize 'head' below 'body' by a 25 pixel high vertical element centered at the bottom. Each structur-


(a)

(b)

(c)

Fig. 13 Part-based articulated objects such as animals or humans. **a** Original images, **b** Color-based segmentation (solution of Eq. (2)). We obtain improved segmentation results **(c)** for further articulated objects based on the novel midrange geometric constraints, e.g., we penalize 'feet' close to 'beak' or 'shoes' above 'hair'. The *bottom left corner* of each image shows the dice-score (and the improvement over the color-based segmentation)

ing element is selected such that it reflects the common label proximities of the specific dataset. 'arms', e.g., mostly appear closer to 'trousers' than 'hands' next to 'hair'. Thus, the structuring elements are chosen such that 'hands' and 'hair' are penalized within a larger distance ($d = 25$) than the labels 'arms' and 'trousers' ($d = 15$).

For the experiments on the Penn-Fudan dataset (Fig. 14) we used the learning approach introduced in Sect. 3.3 and obtained the penalty matrix A and structuring elements $\mathcal{S}_i$ shown in Fig. 5. For example, we penalize the label 'shoes' appearing closely (within 50 pixels) below 'hair' and the label 'face' appearing closely above 'lower clothes'. Figure 14 shows that the proposed constraints improve the semantic labeling of the images compared to (c) the pixel-based approach by Ladicky et al. (2010), (d) the approach by Bo and Fowlkes (2011) who provided the ground truth annotations and (e) the color-based segmentation (solution of Eq. (2)). In the top row, e.g., the incorrect label transition from 'face' to 'lower clothes' is penalized with the novel constraints and the correct label 'upper clothes' is selected.

To allow for a quantitative analysis, we provide the dice-scores (and the improvement over the color-based segmentation) in the bottom left corner of each image. The dice-score (Dice 1945) is given as

$$\frac{2 \cdot \text{True Positives} \cdot 100}{2 \cdot \text{True Positives} + \text{False Negatives} + \text{False Positives}}. \tag{22}$$

Since no multi-label ground truth segmentations are available for the iCoseg and People datasets, we therefore created accurate ground truth labelings (compare Fig. 1d). The qualitative results show improvements up to 12 % of the novel constraints over the color-based segmentation. The novel midrange geometric priors capture richer semantic information and thus allow for a correct semantic interpretation. A discussion of the quantitative results on the Penn-Fudan dataset will be given in Sect. 5.6.

5.2 Part-Based Rigid Objects

An obvious application of the proposed priors are rigid objects consisting of several parts, which is often the case for man-made objects such as cars or bicycles. Using the proposed framework we can improve segmentation results of these objects with all their parts by integrating the proposed

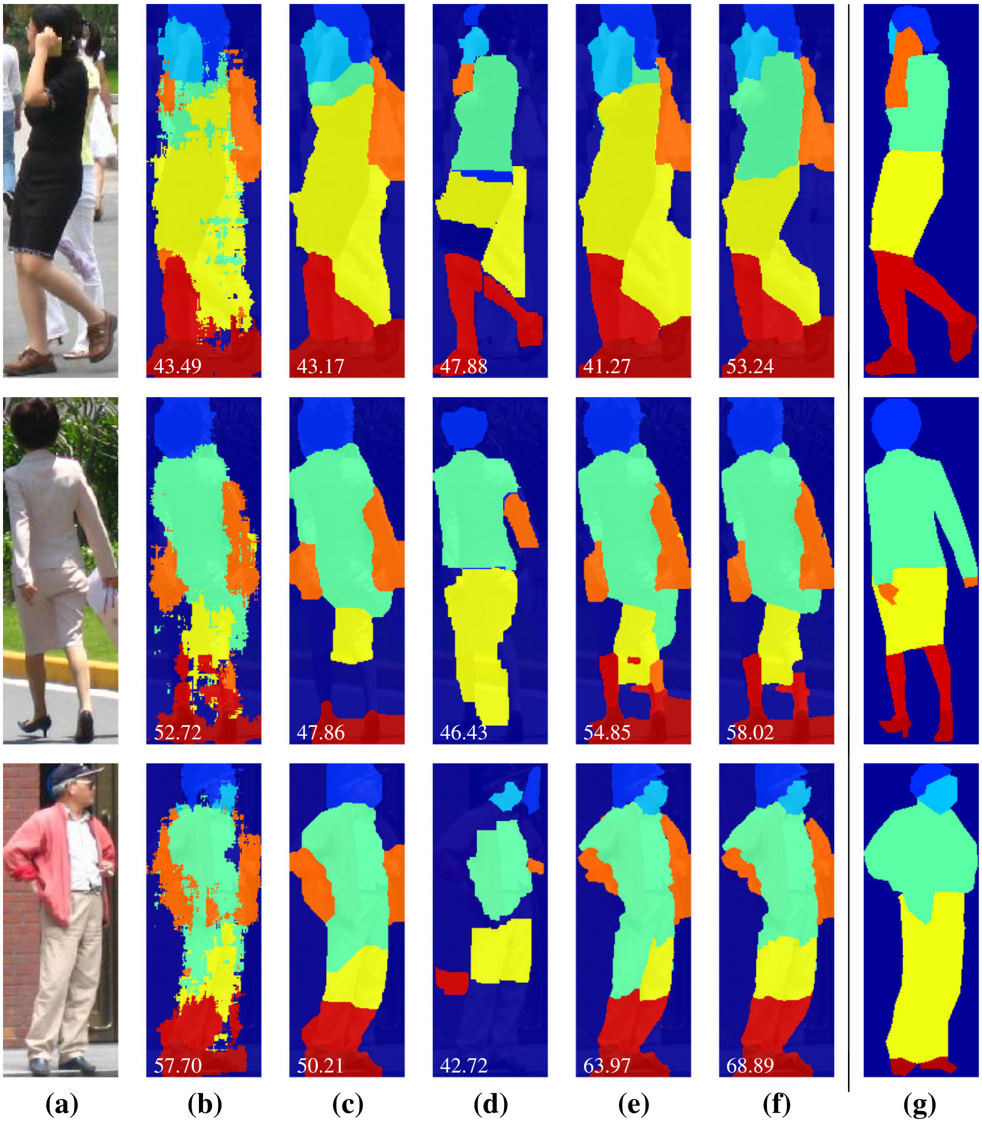


Fig. 14 Improved results on the Penn-Fudan dataset using the learned penalty matrix A and structuring elements $\mathcal{S}_i$ shown in Fig. 5. **a** Original image, **b** Index minimizing (6), **c** Ladicky et al. (2010) pixel-based, **d** Bo and Fowlkes (2011), **e** Solution of Eq. (2), **f** Proposed priors, **g** Ground truth. The proposed novel midrange geometric constraints allow to obtain improved segmentation results by capturing richer semantic information on spatial object inter-relations of part-based articulated objects such as humans

priors. Figure 15 shows results for a set of part-based rigid objects. For example we penalize 'headlight' and 'window' next to each other and 'tires' next to 'headlight' by using 41×41 sparse symmetric elements $\mathcal{S}_i$ (compare Fig. 3d) for $d = 20$). The dice-scores (cf. Eq. (22)) show a significant improvement of more than 6 % compared to the color-based segmentation.

5.3 Scene Segmentation

The proposed constraints are not only useful for part-based objects but can as well be applied to scene segmentation. The same geometric rules that apply to object parts also apply to whole objects within scenes, for example we know that the sky is above the ground and that sheep do not appear close to wolves. In the following, we show results for a variety of scenes in the MSRC benchmark, for the task of facade recognition on the eTRIMS dataset (Korc and Förstner 2009) and for the task of geometric scene labeling of indoor images (Liu et al. 2010).

5.3.1 MSRC Scene Segmentation

In Fig. 16 we show several results from the MSRC benchmark. We compare our results to previous approaches, which incorporate semantic constraints. The global co-occurrence priors by Ladicky et al. (2010) penalize the simultaneous occurrence of specific label sets, but they exhibit two draw-

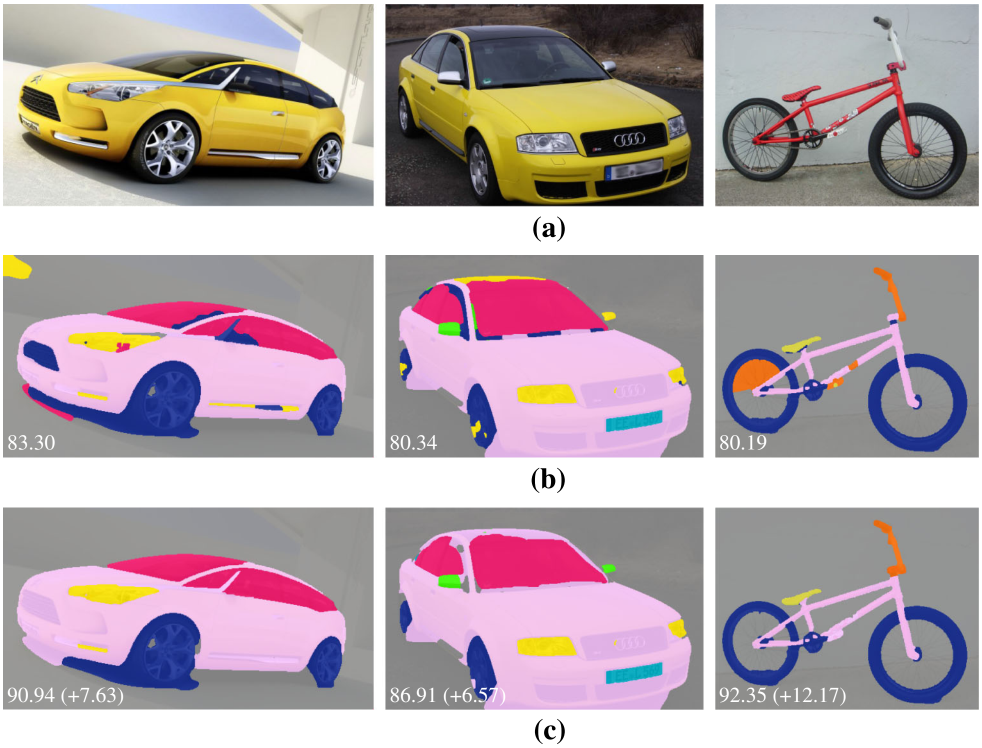


Fig. 15 Part-based rigid objects such as cars or bicycles. **a** Original images, **b** Color-based segmentation (solution of Eq. (2)). We obtain improved segmentation results by imposing the novel geometric priors **(c)**. For example we penalize 'tires' above 'window' or 'handlebar'close to 'tires'

backs: a) The quality of the results depends on the quality of the superpixel partition, which is done prior to any segmentation. This can lead to segmentations such as the cat in Fig. 16 fifth row, where only the black image parts are considered as 'cat'. b) They altogether disregard spatial information. Since the penalty is independent of the size of the regions and their location in the image, the prior is sometimes not strong enough to prevent incorrect label combinations. As a consequence, if more pixels vote for a certain label then they may easily overrule penalties imposed by the co-occurrence term. This can lead to segmentations such as the sheep with cow head (see Fig. 16 first row) despite a large co-occurrence cost for 'sheep' and 'cow'. Other examples are the sign above the book (third row) or the cat below the water (seventh row) despite large costs for 'sign' and 'book' or 'water' and 'cat'.

The local non-metric prior by Strekalovskiy et al. (2012) can be understood as a purely local co-occurrence prior since it only considers directly adjacent pixels as close. If two sheep are standing further apart as in the second row then this case is not penalized by the prior, which can lead to a sheep and a cow close to each other. Besides, this method can easily produce ghost regions, see Sect. 5.5.

There is no notion of distance, direction or proximity in each of the approaches (Ladicky et al. 2010; Strekalovskiy et al. 2012). In contrast, the proposed label cost penalty is proportional to the size of the labeled regions and also effects object labels at larger spatial distances. Hence, the proposed priors are more flexible and allow for the integration of more specific information, which improves segmentation results as shown in the last column (e) of Fig. 16. The result of the cat (see Fig. 16 fifth row), e.g., shows that we can avoid problems which appear due to prior superpixel segmentations.

5.3.2 Facade Parsing on the eTRIMS Dataset

We applied our method for the recognition of facades on the 8-class eTRIMS facade dataset (Korc and Förstner 2009). The following eight object classes are considered: 'sky', 'building', 'window', 'door', 'vegetation', 'car', 'road' and 'pavement'.

In Fig. 17 we present five examples of facade segmentations. In columns one and two, the incorrect label transition from 'window' (blue) to 'door' (yellow) is corrected with the novel constraints by penalizing the appearance of 'window' close to 'door'. In columns three and four, the wrong labeled 'sky' pixels (cyan) in the middle of the building disappear by claiming that no other region appears above 'sky'. The combination of both constraints improves the segmentation in the rightmost column, where both the incorrect 'sky' and the incorrect 'door' pixels are removed with the novel constraints.

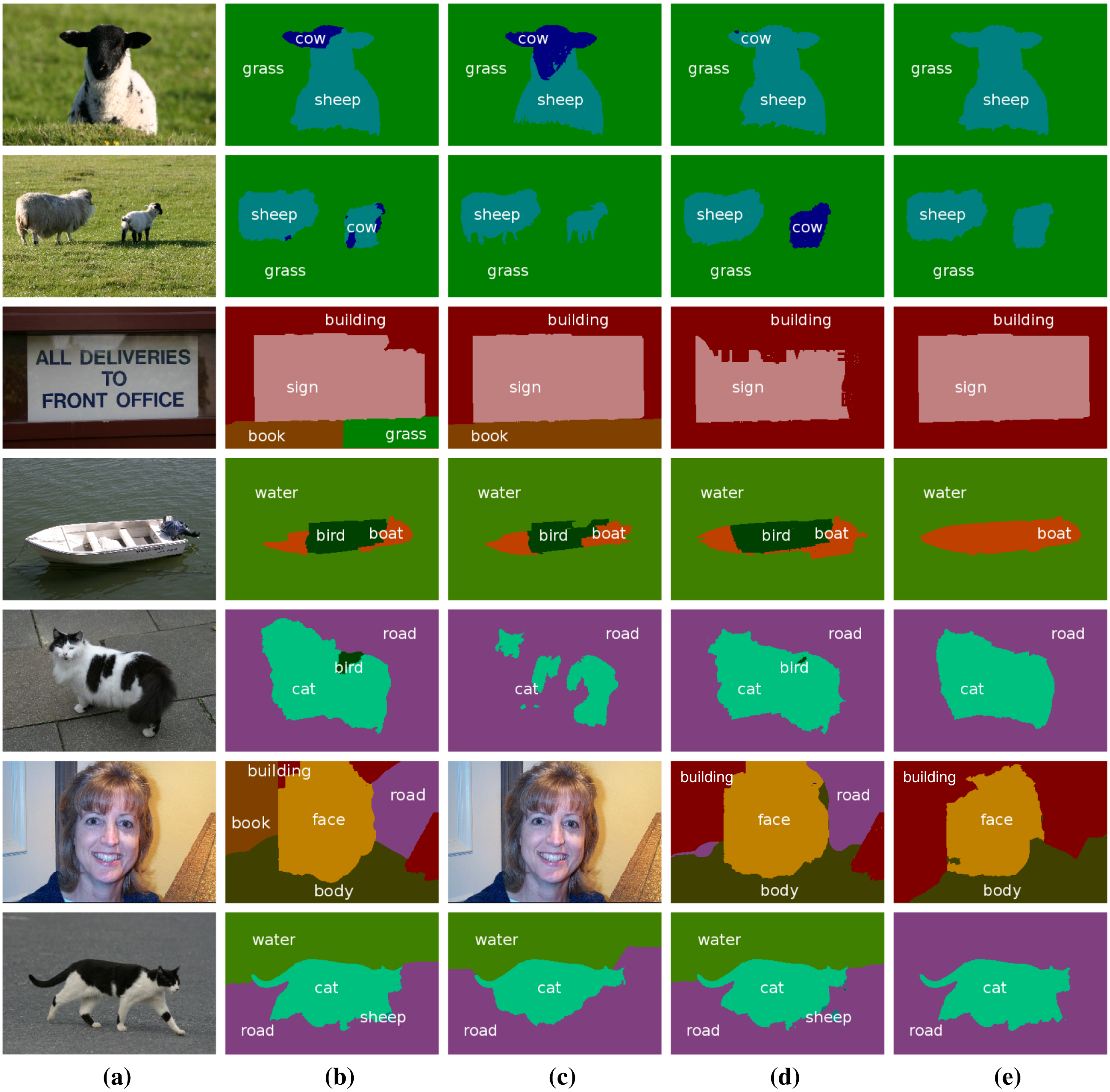


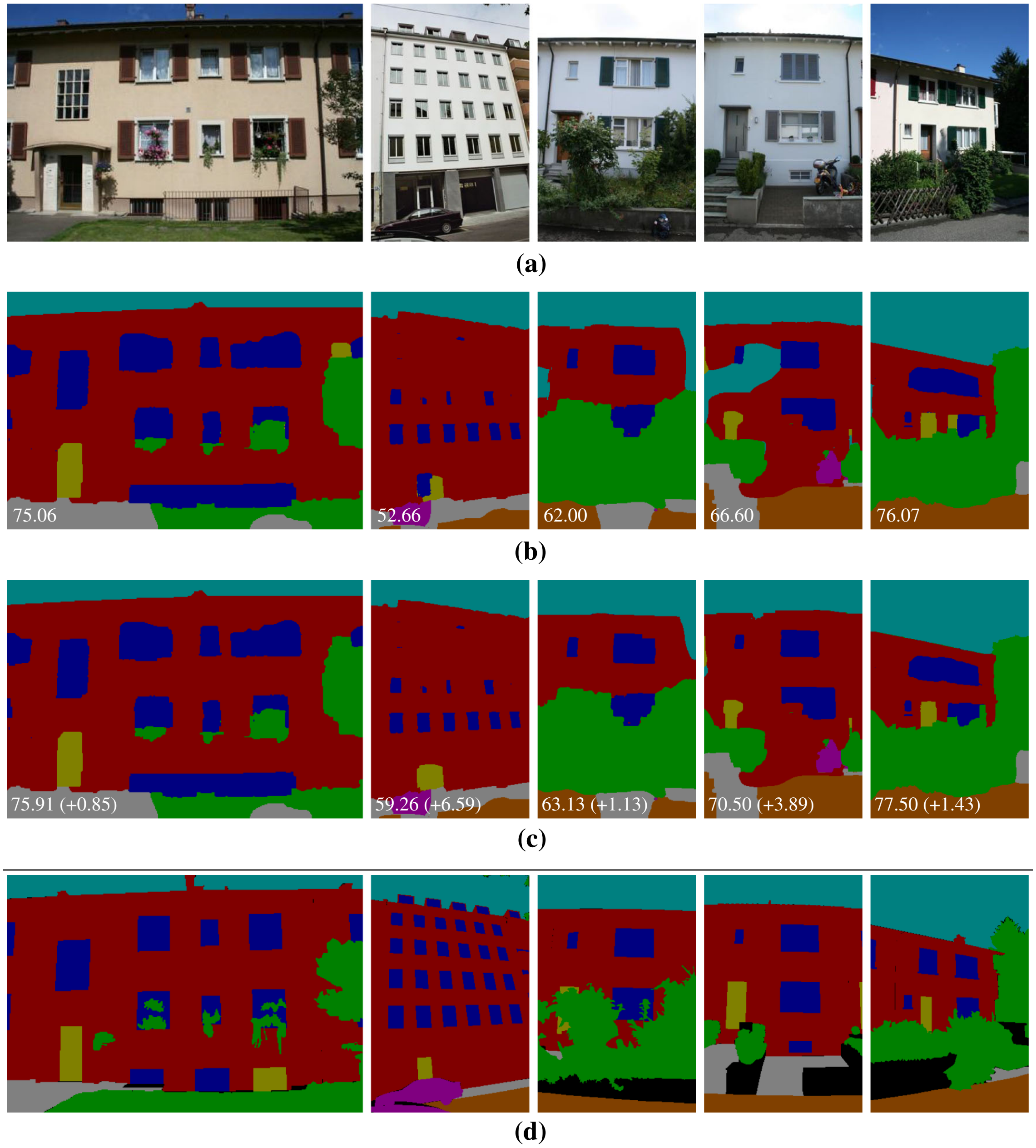
Fig. 16 Improved results on the MSRC benchmark. **a** Original image, **b** Solution of Eq. (2), **c** Global co-occurrence prior (Ladicky et al. 2010), **d** Local non-metric prior (Strekalovskiy et al. 2012), **e** Proposed geometric priors. Midrange geometric priors capture richer semantic information on spatial object inter-relations such as distances, direction and relative location than previous approaches such as global co-occurrence (Ladicky et al. 2010) or local co-occurrence (Strekalovskiy et al. 2012)

A first quantitative comparison is provided by the dice-scores. A concrete benchmark analysis will be given in Sect. 5.6.

5.3.3 *Geometric Class Labeling of Indoor Images*

In tasks like 3D reconstruction or vision-guided robot navigation a rough labeling of the environment is essential. In particular, the geometric classes such as 'floor' or 'right wall' are of importance. We therefore applied our novel constraints on the dataset of indoor images from Liu et al. (2010) with the five-regions layout: 'left wall' (yellow), 'floor' (green), 'right wall' (pink), 'ceiling' (blue) and 'center' (cyan).

Knowing, for example, that except the ceiling no other region appears above the left wall, the incorrect labels within the region 'left wall' can be removed. The midrange geometric constraints can, e.g., be defined such that they penalize everything above 'ceiling' and everything above 'left'/'right'/'center' except 'ceiling'. Results for six different images with the corresponding dice-scores are shown in


(a)

(b)

(c)

(d)

Fig. 17 Improved labeling of facades on the eTRIMS benchmark. **a** Original image, **b** Color-based segmentation (solution of Eq. (2)), **c** Segmentation with novel constraints, **d** Ground truth. By penalizing 'window' (*blue*) close to 'door' (*yellow*) and by claiming that no other region appears above 'sky' (*cyan*) the incorrectly labeled 'door', 'window' and 'sky' pixels disappear

Fig. 18. A quantitative benchmark analysis will be given in Sect. 5.6.

5.4 Analysis of Failure Cases

In order to evaluate the strengths and weaknesses of our approach we looked into a number of failure cases on the MSRC benchmark and compared our results to the index minimizing the appearance model (6) and the results by Ladicky et al. (2010) and Strekalovskiy et al. (2012), see Fig. 19 for some examples. After close investigation of many cases we can formulate one main reason for incorrect labelings:

The appearance term (see Sect. 2.2; Eq. (6)) used by all three approaches favors incorrect labels over the correct one (Fig. 19c). Take for example the 'building' which occurs in the first row in all three results instead of the correct label 'boat'. Since the appearance term clearly favors the white color to belong to a 'building' and 'building' and 'water' is not an uncommon combination in the penalty matrix we

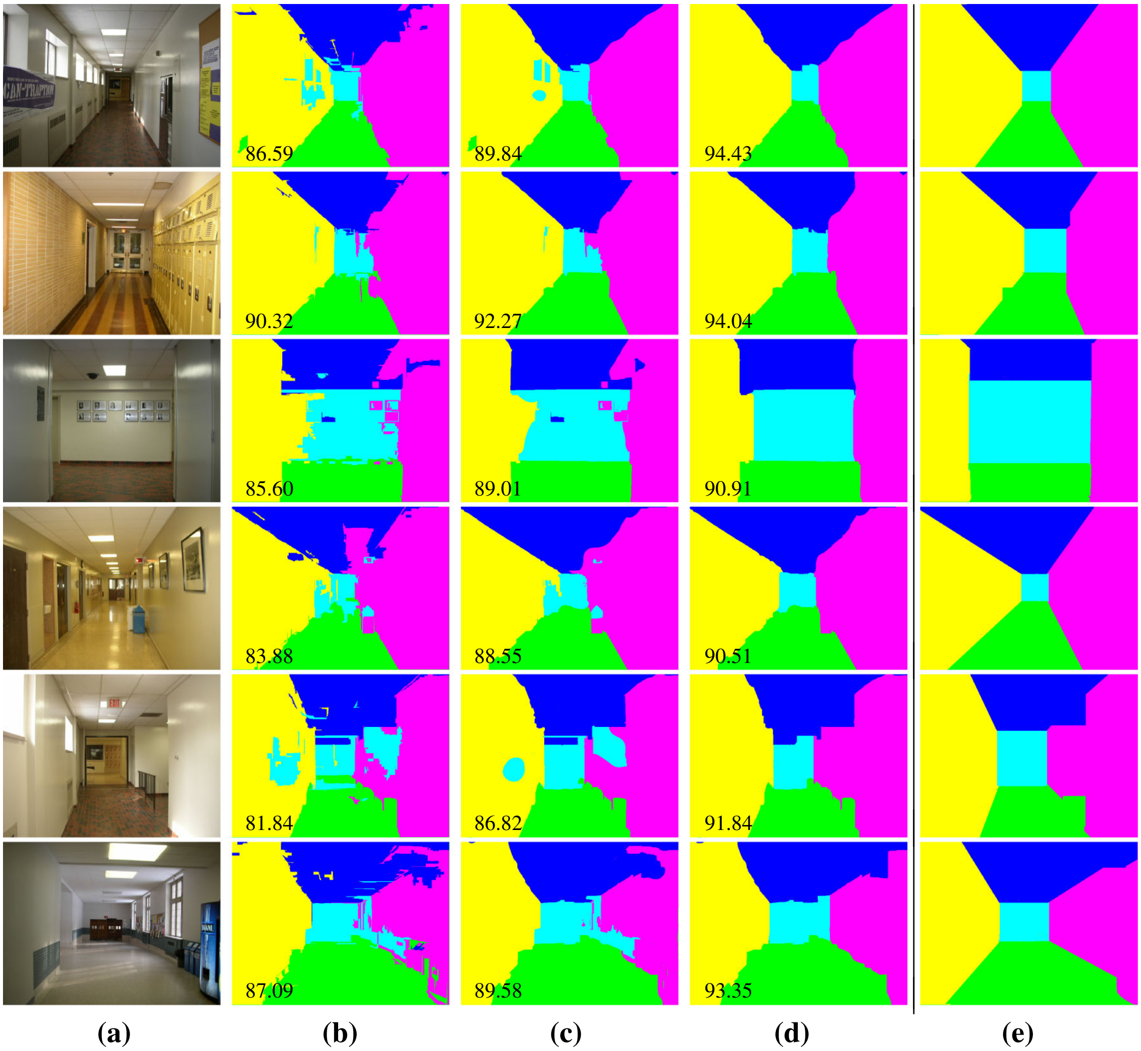


Fig. 18 Corrected layout of labels. **a** Original image, **b** Index minimizing (6), **c** Solution of Eq. (2), **d** Proposed geometric priors, **e** Ground truth segmentation. The novel priors allow a correct segmentation of the corridors by including directional relations such as that the floor usually is below the ceiling

obtain incorrect labels. The same happens for the examples in the central and bottom row, where the appearance term yields lots of incorrect labels. Since the appearance term favors 'building' over 'sign' in the bottom row (see column c) and the proposed priors do not favor 'sign' close to 'sky' over 'building', the incorrect segmentation results. This happens in a similar way in the central row, where the appearance term suggests 'car' next to 'road' and 'water'. Since 'car' is more likely to occur above 'road' than 'water' the water is assigned the label 'road'.

Even though none of the methods yields good results for these images, the proposed novel constraints at least yield a reasonable combination of labels in contrast to the other methods. These failure cases suggest that improvements of the method can be gained by using better appearance models.

5.5 Preventing Ghost Labels

'Ghost labels' denote thin artificial regions which are easily introduced if label distances are learned from training data, see for example Strekalovskiy et al. (2012). If the distance function, i.e. the penalty matrix A, does not obey the triangle inequality 'ghost labels' can appear. They reduce costs of direct label transitions by taking a 'detour' over a third, unrelated but less expensive label. For example, the labels sheep and grass are common next to each other, and the same holds for cow and grass, but cows usually do not occur directly next to sheep, so the triangle inequality is violated.

Examples are given in Fig. 20b with a close-up in Fig. 20c. The segmentation result obtained by Strekalovskiy et al. (2012), e.g., contains very thin 'boat' regions at the edge

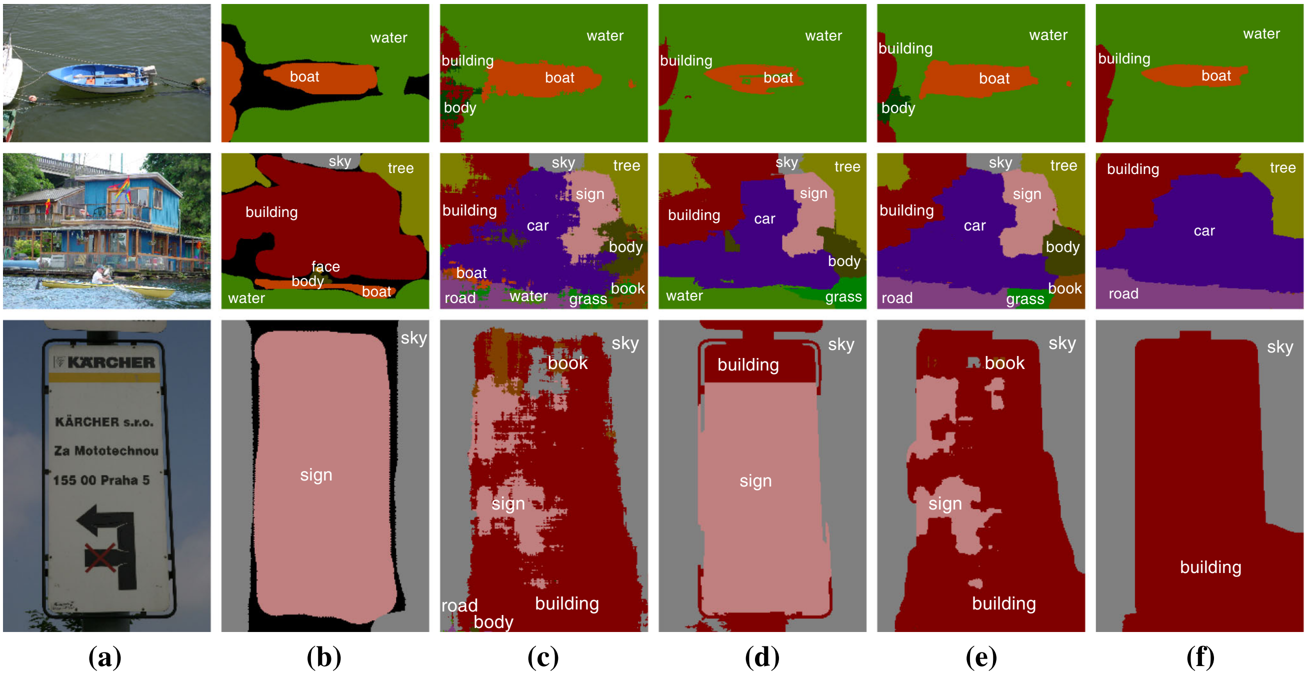


Fig. 19 Analysis of failure cases. **a** Original image, **b** Ground truth, **c** Index minimizing (6), **d** Global co-occurrence prior (Ladicky et al. 2010), **e** Local non-metric prior (Strekalovskiy et al. 2012), **f** Proposed geometric priors. For a thorough evaluation we looked into the failure cases of our approach and compared to the index minimizing the appearance term (6) and the results of Ladicky et al. (2010) and Strekalovskiy et al. (2012). We identified one main reason: the appearance term favors incorrect labels

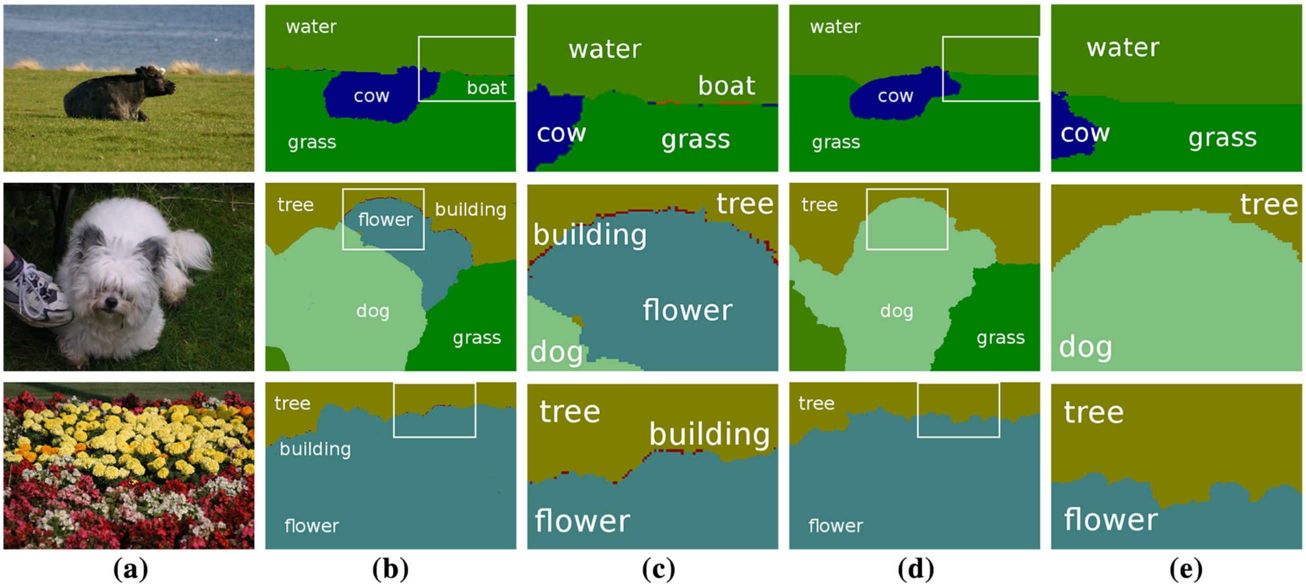


Fig. 20 Midrange geometric priors prevent ghost labels. **a** Original image, **b** Local prior (Strekalovskiy et al. 2012), **c** Zoom of **(b)**, **d** Novel geometric prior, **e** Zoom of **(d)**. If the transition of two labels is cheaper via a third label artificial labels will be introduced as shown in **(b)** and as close-up in **(c)**. The proposed geometric priors consider regions with more than one-pixel distance still as adjacent and thus avoid ghost labels

of the ‘grass’ label, because the transition between the labels ‘water’ and ‘boat’ and ‘boat’ and ‘grass’ is in sum less costly than the direct transition between ‘water’ and ‘grass’. The computed label distance matrix denotes the following distances (Strekalovskiy et al. 2012):

$$\begin{aligned} d(\text{‘grass’}, \text{‘water’}) = 7.0 > 4.7 = d(\text{‘grass’}, \text{‘boat’}) \\ + d(\text{‘boat’}, \text{‘water’}), \end{aligned}$$

thus, the more costly label transitions from ‘grass’ to ‘water’ is avoided by introducing infinitesimal ‘boat’ regions.

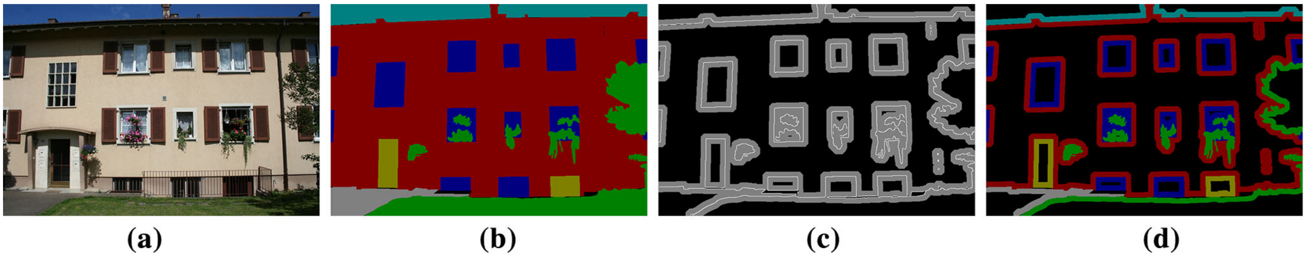


Fig. 21 Ground truth and trimap segmentations. **a** Original image, **b** Ground truth segmentation, **c** Trimap of **(b)**, **d** Trimap segmentation of **(b)**. We evaluate the performance using different evaluation domains: **b** The whole image domain; **c**, **d** Trimap of (**b**) generated by taking a 13 pixel band surrounding the object boundaries

The proposed geometric priors prevent ghost regions since the size of the structuring element is usually larger than two pixels and thus considers more than a single pixel wide margin as close to the object. This leads to overlaps in indicator functions which are larger than a single pixel and thus much more expensive than in the approach by Strekalovskiy et al. (2012), see for example our results in Fig. 20d with a close-up in Fig. 20e. Our learned penalization matrix A, e.g., indicates the following penalties:

$$A(\text{‘grass’}, \text{‘water’}) = 6.2 < 9.6 = A(\text{‘grass’}, \text{‘boat’}) \\ +A(\text{‘boat’}, \text{‘water’}).$$

Thus, the direct transition from ‘grass’ to ‘water’ is favored in the optimization process.

5.6 Benchmark Evaluation

In the following we will show quantitative results on the aforementioned benchmarks and compare our segmentations to state-of-the-art approaches for semantic labeling. For the benchmark analysis, we computed three different evaluation scores. The scores denote the average accuracy on the benchmark given as

$$\frac{\text{True Positives} \cdot 100}{\text{True Positives} + \text{False Negatives}}$$

per pixel and per class and the dice-score averaged over all images. The dice-score (Dice 1945) additionally takes the false positives into account and is given in Eq. (22).

We measure the labeling accuracies using the different evaluation scores and using different evaluation regions. The evaluation region can be the whole image domain or restricted to a band surrounding the region boundaries. The restricted evaluation regions are called *trimap* (Kohli et al. 2009). An exemplary trimap with an evaluation band width of 13 is illustrated in gray in Fig. 21c.

5.6.1 Penn-Fudan Benchmark Scores

The Penn-Fudan pedestrian benchmark (Wang et al. 2007) includes 169 images with an average resolution of 290 × 116 pixels and 12 different labels such as ‘hair’, ‘face’, ‘left leg’ or ‘right leg’. We follow Bo and Fowlkes (2011) and combine the left and right hand/leg/shoe to one region each, resulting in the 8 different labels: ‘hair’, ‘face’, ‘upper clothes’, ‘lower clothes’, ‘arm’, ‘leg’, ‘shoes’, ‘background’.[2] For the benchmark experiments, we divided the image set randomly into 60 % training and 40 % test images and learned the penalty matrix A and structuring elements $\mathcal{S}_i$ (see Fig. 5) as described above in Sect. 3.3. The parameter lambda is set to $\lambda = 0.8$.

In Tables 1 and 2 we compare the performance of our method with the approaches by Ladicky et al. (2010) for the pixel-based prior, Bo and Fowlkes (2011) with the shape-based model and the recently proposed work of Luo et al. (2013) for pedestrian parsing. Furthermore, we present the accuracy of the index minimizing the appearance model (6) and the solution of the approach without geometric priors, i.e. the solution of Eq. (2). We evaluate the performance of the approaches on the whole image domain and on the trimap with band width 13. Table 1 shows that the proposed midrange geometric constraints outperform the related state-of-the-art segmentation algorithms. In Table 2 we additionally compare the confusion matrices on both evaluation domains. Green colored values indicate that the proposed method outperforms the comparative approach for this region. The proposed priors achieve the best performance for the vast majority of regions.

5.6.2 MSRC Segmentation Benchmark Results

In the following we will show quantitative results on the MSRC database and compare our segmentations to state-of-the-art approaches for semantic labeling.

The MSRC benchmark comprises 591 images with a resolution of 320 × 213 pixels which contain 21 different labels

[2] Note that (Bo and Fowlkes 2011) additionally neglected the region ‘shoes’.

Table 1 Penn–Fudan benchmark scores

	Evaluation on the whole image domain			Evaluation on the trimap (width 13)		
	Accur. per pixel	Accur. per class	Dice-score	Accur. per pixel	Accur. per class	Dice-score
Index minimizing (6)	66.97	67.81	55.63	58.09	64.28	53.28
Solution of Eq. (2)	72.83	70.61	59.98	65.93	68.03	58.58
Ladicky et al. (2010) pixel-based	71.84	67.36	57.21	64.62	64.70	55.52
Bo and Fowlkes (2011)	–	57.29	–	–	–	–
Luo et al. (2013)	–	54.7	–	–	–	–
Proposed midrange geometric priors	**73.84**	**70.78**	**60.65**	**67.00**	**68.26**	**59.15**

The proposed constraints outperform the related state-of-the-art segmentation algorithms on the Penn–Fudan benchmark
The best results are given in bold

Table 2 Confusion matrix on the Penn–Fudan dataset obtained for the evaluation on the whole image domain and on the trimap

	Evaluation on the whole image domain								Evaluation on the trimap (width 13)							
	Backgr.	Hair	Face	Up. Cl.	Low. Cl.	Arms	Legs	Shoes	Backgr.	Hair	Face	Up. Cl.	Low. Cl.	Arms	Legs	Shoes
Background	**72**	3	1	6	9	3	1	5	**59**	5	2	8	13	5	1	8
Hair	4	**81**	6	8	0	1	0	0	4	**81**	6	8	0	1	0	0
Face	4	19	**71**	34	0	2	0	0	4	19	**71**	4	0	2	0	0
Upper Clothes	9	1	1	**78**	5	6	0	0	10	1	2	**73**	6	7	0	0
Lower Clothes	7	0	0	7	**78**	2	1	5	8	0	0	7	**76**	2	1	6
Arms	16	0	4	23	5	**53**	0	0	16	0	4	23	5	**53**	0	0
Legs	19	0	0	0	22	3	**51**	6	19	0	0	0	22	3	**51**	6
Shoes	9	0	0	0	5	0	3	**83**	9	0	0	0	5	0	3	**83**
Index minimizing (6)	8	3	1	9	5	–4	1	0	13	4	1	12	6	–4	1	0
Solution of Eq. (2)	2	0	1	2	0	–2	0	–1	2	0	1	2	0	–2	0	–1
Ladicky et al. (2010) pixel-based	3	1	5	1	0	3	12	2	4	2	5	0	0	3	12	2
Bo and Fowlkes (2011)	–9	36	10	3	7	27	9	–	–	–	–	–	–	–	–	–
Luo et al. (2013)	–13	36	17	0	3	28	1	–	–	–	–	–	–	–	–	–

The elements (i, j) represent the percentage of pixels labeled i by the method and j in the ground truth. We compare the difference between our method and the comparison ones along the diagonal (shown in bold). The values are given in green when the proposed method outperforms the comparative approach, in red otherwise

such as ‘cow’, ‘book’, ‘building’ or ‘grass’. To conduct experiments on this benchmark, we follow Ladicky et al. (2010) and divide the image set randomly into 60 % training and 40 % test images. For the benchmark experiments we chose a symmetric set $\mathcal{S}$ of size 9×9 for all labels (compare Fig. 3d) and selected $\lambda = 0.3$. The proximity matrix A is learned on the training set as described above in Sect. 3.4 and illustrated in Fig. 6.

To evaluate the segmentation accuracy of the proposed method, in Table 3 we compare the benchmark scores of our method to state-of-the-art segmentation algorithms with co-occurrence priors: the approaches by Gould et al. (2008) with relative location priors, Ladicky et al. (2010) for the pixel-based and the co-occurrence and hierarchical prior, Lucchi et al. (2011) for the data pairwise global and local models, Vezhnevets et al. (2011) for the weakly and fully supervised approach and Strekalovskiy et al. (2012) with the non-metric distance functions for multi-label problems. Moreover, we present the accuracy of the index minimizing the appearance model (6) and the solution of Equation (2). The results indicate that we outperform the other co-occurrence based methods in average class and pixel accuracy.

Note that the high score of the approach by Strekalovskiy et al. (2012) does not reflect the ghost label problem since a) these regions contain only very few pixels, and b) these pixels occur in mostly unlabeled regions of the ground truth near object boundaries, see the second column in Fig. 19. However, the introduction of entirely unrelated objects, albeit small ones, is often problematic for applications.

The benchmark results in general suggest rather small improvements for the integration of geometric spatial priors. This is somewhat surprising since the images show strong improvements and the prior corresponds to typical human thinking. As already mentioned by Lucchi et al. (2011) who

Table 3 MSRC benchmark scores. We compare the segmentation accuracy to state-of-the-art segmentation algorithms with co-occurrence priors on the MSRC benchmark. The approach by Ladicky et al. (2010) in the last row is added for the sake of completeness. They use a more sophisticated appearance model and model co-occurrence by an additional cost function which can be seen as potentials of the highest order $|\Omega|$, instead of order two as in our approach[a]

	Accur. per pixel	Accur. per class	Dice-score	Building	Grass	Tree	Cow	Sheep	Sky	Plane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
Gould et al. (2008) CRF + rel. loc.	76.5	64.38	–	**72**	95	81	66	71	93	74	70	70	69	72	68	55	23	82	40	77	60	50	50	14
Ladicky et al. (2010) pixel-based	84	77.72	81	69	**97**	91	86	88	96	88	82	91	82	93	82	63	44	92	66	86	80	56	73	30
Lucchi et al. (2011), DPG local	75	68.62	–	54	88	83	79	82	95	87	70	85	81	**97**	69	**72**	27	88	46	60	74	27	49	28
Lucchi et al. (2011), DPG loc.+glob.	80	74.62	–	65	87	87	84	75	93	**94**	78	83	72	93	86	70	50	**93**	**80**	**86**	78	28	58	27
Vezhnevets et al. (2011), weak sup.	67	66.52	–	12	83	70	81	**93**	84	91	55	**97**	87	92	82	69	**51**	61	59	66	53	44	9	**58**
Vezhnevets et al. (2011), full sup.	72	71.71	–	21	93	77	86	**93**	96	92	61	79	**89**	89	**89**	68	50	74	54	76	68	47	49	55
Strekalovskiy et al. (2012)	84.85	77.49	80.79	70	**97**	**92**	**89**	85	96	81	**83**	90	82	92	83	66	45	92	63	**86**	80	51	73	32
Index minimizing (6)	82.58	75.91	79.51	67	**97**	91	85	86	95	88	81	90	82	93	81	62	43	65	66	**86**	79	54	73	31
Solution of Eq. (2)	82.98	76.28	79.66	68	96	91	86	87	95	88	81	90	82	93	81	62	43	65	66	**86**	80	55	72	30
Proposed midrange geometric priors	**84.97**	**78.19**	**81.04**	69	**97**	**92**	87	87	**97**	87	82	91	83	94	84	62	44	**93**	67	**86**	**83**	**57**	**74**	26
Ladicky et al. (2010) hier. + co-oc.	86.76	76.76	80.78	82	95	88	73	88	100	83	92	88	87	88	96	96	27	85	37	93	49	80	65	20

The best results are given in bold

[a] One could argue that with the introduction of auxiliary variables the model (Ladicky et al. 2010) can be reduced to a model of order two. However, the nature of the problem stays a problem of higher order with respect to the original variables. In contrast, our model is of order two without auxiliary variables

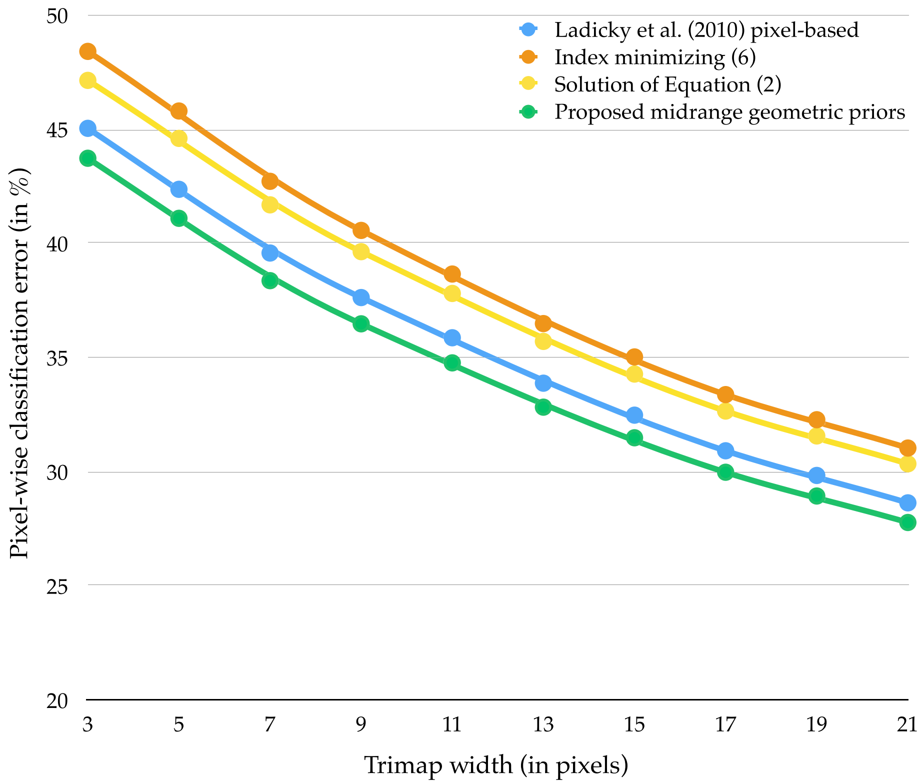


Fig. 22 Pixel-wise classification error on the MSRC benchmark. With increasing width of the evaluation region, the pixel-wise classification error decreases. The best classification is achieved with the proposed midrange geometric priors. For the computation of the trimaps, the more precise ground truth labeling of Malisiewicz and Efros (2007) has been used

stated similar findings this is probably due to the rather crude ground truth of the benchmark with large unlabeled regions especially at object boundaries, compare Fig. 19b. These regions are not counted in the score, but nevertheless leave a lot of room for misclassification or improvements. Therefore, we think that the benchmark score should not be overstressed here.

To provide a second evaluation measure, we additionally computed the classification error on the precise ground truth provided by Malisiewicz and Efros (2007). In Fig. 22 we compare the pixel-wise classification error for different widths of the evaluation region. We consider trimaps with 3 to 21 pixels wide bands surrounding the object boundaries (cf. Fig. 21). The classification error decreases with increasing width of the trimap. The smallest error is achieved with the proposed midrange geometric priors.

Qualitative comparisons with the two best scoring of the above mentioned methods by Ladicky et al. (2010) with co-occurrence and hierarchical prior and by Strekalovskiy et al. (2012) on the MSRC database are given in Fig. 16. The results show that the proposed method reduces the number of mislabeled objects.

5.6.3 *eTRIMS Facade Parsing Benchmark Results*

In Sect. 5.3 we already demonstrated some qualitative results for the task of segmenting facades. The 8-class eTRIMS facade dataset (Korc and Förstner 2009) consists of 60 images with a resolution of 512 × 768. Again, we split the dataset into 60/40 for training and testing and set $\lambda = 0.6$.

In Table 4 we compare the accuracy per pixel on the whole image domain as well as for different band widths of the trimap. For all evaluation domains the best score is achieved with the proposed priors. The relatively minor improvement in the percentages reflects our observation that significant improvements of the semantic segmentation do not necessarily lead to a substantial improvement of the score. In Fig. 17 4th column, e.g., a major part of the image—namely the mislabeled 'sky' pixels—is corrected by the proposed constraints. The dice-score for this image, however, only improved by 3.9 %.

5.6.4 *Score for the Task of Geometric Class Labeling of Indoor Images*

The definition of the geometric classes in a scene is another interesting application area. For our experiments we use the indoor dataset from Liu et al. (2010) which consists of 300 indoor images with a resolution of 640 × 480 pixels. To guarantee comparability we use their appearance model and set $\lambda = 1$.

In Table 5 we compare our results to the approaches by Liu et al. (2010) and Strekalovskiy and Cremers (2011) who use the same appearance model. We achieved an overall accuracy of 87.24 %, compared Liu et al. with 85 % and Strekalovskiy and Cremers with 85.3 %.

Table 4 The highest scores on the eTRIMS benchmark are achieved with the proposed priors

	Trimap width 9	Trimap width 13	Trimap width 17	Trimap width 21	Accur. per pixel
Index minimizing (6)	63.09	67.90	71.28	73.58	80.56
Solution of Eq. (2)	69.33	73.35	76.17	78.08	84.36
Ladicky et al. (2010) pixel-based	68.79	72.84	75.75	77.76	84.22
Proposed midrange geometric priors	**69.37**	**73.46**	**76.34**	**78.31**	**84.82**

The scores are the accuracies per pixel computed on different trimap segmentations and the whole image domain
The best results are given in bold

Table 5 Improved score for the task of geometric class labeling

	Accur. per pixel	Accur. per class	Dice- score
Index minimizing (6)	84.99	79.97	77.67
Solution of Eq. (2)	86.64	81.59	79.51
Liu et al. (2010)	85	–	–
Strekalovskiy and Cremers (2011)	85.3	–	–
Proposed midrange geometric priors	**87.24**	**81.90**	**80.17**

The proposed midrange geometric constraints outperform the approaches by Liu et al. (2010) and Strekalovskiy and Cremers (2011) which use the same appearance model
The best results are given in bold

Table 6 Average runtimes for multi-label segmentation of an image of the iCoseg (Batra et al. 2010) and the People (Ramanan 2006) dataset containing 4–9 labels

	Average runtime
Without geometric prior	2.29 s
With geometric prior	7.74 s

5.7 Runtimes

We finally investigate the runtime of the proposed method.

Apart from the size of the $\mathcal{S}_i$, the runtime mainly depends on the number of labels used for the segmentation. For the MSRC benchmark 21 labels have been used. Usually, images consist of less than ten different labels, e.g., images of persons can include hair, head, body, arms, hands, trousers, legs, shoes or background.

We obtain average runtimes of 7.7 s on the iCoseg (Batra et al. 2010) and the People (Ramanan 2006) dataset (see Table 6) compared to 2.3 s if we do not use the novel priors. The images have a resolution of around 500 × 333 pixels and the sets $\mathcal{S}_i$ have a size of around $d = 25$.

The MSRC benchmark, in contrast, contains 21 labels, which in theory can appear all at the same time in a single image. This leads to lots of label pairs, most of which are highly unlikely. To reduce the runtime of the approach we used sparse structuring elements $\mathcal{S}_i$ yielding equivalent results to full elements in around 180 s on average (note that

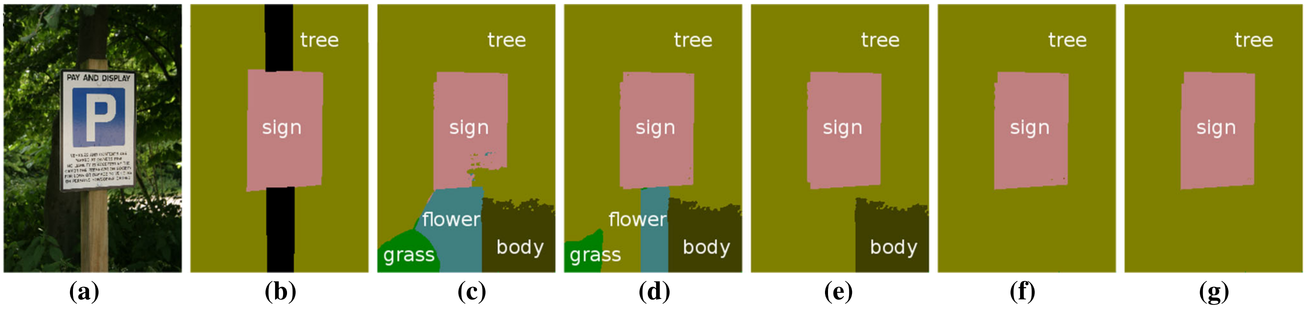


Fig. 23 Minimizing runtime. To minimize runtime in case of large label numbers we use sparse structuring elements (SE). The evolution of the solution for an increasing number of entries in a structuring element $\mathcal{S}$ of size 15×15 shows that very few entries (here 10 entries in a 15×15 SE) are already sufficient to obtain accurate results: **a** Original, **b** Ground truth, **c** $|\mathcal{S}| = 0$, 13 s; **d** $|\mathcal{S}| = 5$, 152 s, **e** $|\mathcal{S}| = 7$, 163 s, **f** $|\mathcal{S}| = 10$, 176 s, **g** $|\mathcal{S}| = 225$, 914 s. The runtimes denote the average runtime on the MSRC benchmark for 21 labels with the respective number of entries $|\mathcal{S}|$

we do not work on superpixels). We can conclude that already very sparse sets $\mathcal{S}_i$ containing around ten entries yield results very similar to the full set $\mathcal{S}_i$ (compare Fig. 23).

6 Conclusion

In this article we introduced a framework for the integration of midrange geometric priors into semantic segmentation and recognition within a variational multi-label approach. Midrange geometric priors impose constraints on directions and/or distances in which label pairs usually occur. We call them midrange, since the constraints are neither global by taking all pixels into consideration such as co-occurrence priors nor are they purely local by only regarding single pixels or pairwise pixel interactions. Instead, the user is able to define the range and specific shape of the interactions between pixels that are penalized. We have shown how morphological operations such as the continuous formulation of the dilation operation can be employed to formulate these constraints within a continuous optimization approach. We gave a convex relaxation, which guarantees independence of the initialization.

The proposed approach does not require the computation of superpixels and prevents the emergence of one pixel wide ‘ghost labels’. Experiments show that the proposed novel constraints are beneficial for many segmentation scenarios, e.g., for part-based articulated objects such as humans, animals or clothes, for part-based rigid objects, especially man-made items, and for semantic scene segmentation.

Acknowledgments We thank three anonymous reviewers for their constructive feedback.

References

Arbelaez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., & Malik, J. (2012). Semantic segmentation using regions and parts. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Batra, D., Kowdle, A., Parikh, D., Luo, J., & Chen, T. (2010). iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bergbauer, J., Nieuwenhuis, C., Souiai, M., & Cremers, D. (2013). Proximity priors for variational semantic segmentation and recognition. In *ICCV Workshop on Graphical Models for Scene Understanding*.

Bo, Y., & Fowlkes, C. C. (2011). Shape-based pedestrian parsing. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Carreira, J., Caseiro, R., Batista, J., & Sminchisescu, C. (2012). Semantic segmentation with second-order pooling. In *European Conference on Computer Vision (ECCV)*.

Carreira, J., & Sminchisescu, C. (2012). CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *34*(7), 1312–1328.

Chambolle, A., & Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision (JMIV)*, *40*(1), 120–145.

Delong, A., & Boykov, Y. (2009). Globally optimal segmentation of multi-region objects. In *IEEE International Conference on Computer Vision (ICCV)*.

Delong, A., Gorelick, L., Veksler, O., & Boykov, Y. (2012). Minimizing energies with hierarchical costs. *International Journal on Computer Vision (IJCV)*, *100*(1), 38–58.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Felzenszwalb, P. F., & Veksler, O. (2010). Tiered scene labeling with dynamic programming. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fröhlich, B., Rodner, E., & Denzler, J. (2012). Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach. In *Asian Conference on Computer Vision (ACCV)*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gould, S., Rodgers, J., Cohen, D., Elidan, G., & Koller, D. (2008). Multiclass segmentation with relative location prior. In *International Journal on Computer Vision (IJCV)*.

Kohli, P., Kumar, M. P., Torr, P. H. S.: P3 & beyond: Solving energies with higher order cliques. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)* (2007).

Kohli, P., Ladicky, L., & Torr, P. H. S. (2009). Robust higher order potentials for enforcing label consistency. *International Journal on Computer Vision (IJCV)*, *82*(3), 302–324.

Komodakis, N., & Paragios, N. (2009). Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kontschieder, P., Kohli, P., Shotton, J., & Criminisi, A. (2013). Geof: Geodesic forests for learning coupled predictors. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Korc, F., & Förstner, W. (2009). *eTRIMS Image Database for interpreting images of man-made scenes*. Technical Report, Department of Photogrammetry, University of Bonn.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2009). Associative hierarchical CRFs for object class image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. H. S. (2010). Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision (ECCV)*.

Liu, X., Veksler, O., & Samarabandu, J. (2010). Order-preserving moves for graph-cut-based optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*(7), 1182–1196.

Lucchi, A., Li, Y., Boix, X., Smith, K., & Fua, P. (2011). Are spatial and global constraints really necessary for segmentation? In *IEEE International Conference on Computer Vision (ICCV)*.

Luo, P., Wang, X., & Tang, X. (2013). Pedestrian parsing via deep decompositional network. In *IEEE International Conference on Computer Vision (ICCV)*.

Malisiewicz, T., Efros, A. A. (2007). Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference (BMVC)*.

Michelot, C. (1986). A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, *50*(1), 195–200.

Möllenhoff, T., Nieuwenhuis, C., Toeppe, E., & Cremers, D. (2013). Efficient convex optimization for minimal partition problems with volume constraints. In *Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*.

Nieuwenhuis, C., & Cremers, D. (2013). Spatially varying color distributions for interactive multi-label segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(5), 1234–1247.

Nieuwenhuis, C., Strekalovskiy, E., & Cremers, D. (2013). Proportion priors for image sequence segmentation. In *IEEE International Conference on Computer Vision (ICCV)*.

Nieuwenhuis, C., Töppe, E., & Cremers, D. (2013). A survey and comparison of discrete and continuous multi-label optimization approaches for the Potts model. *International Journal on Computer Vision (IJCV)*, *104*(3), 223–240.

Nosrati, M., Andrews, S., & Hamarneh, G. (2013). Bounded labeling function for global segmentation of multi-part objects with geometric constraints. In *IEEE International Conference on Computer Vision (ICCV)*.

Pock, T., & Chambolle, A. (2011). Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *IEEE International Conference on Computer Vision (ICCV)*.

Pock, T., Chambolle, A., Bischof, H., & Cremers, D. (2009). A convex relaxation approach for computing minimal partitions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pock, T., Cremers, D., Bischof, H., & Chambolle, A. (2009). An algorithm for minimizing the Mumford–Shah functional. In *IEEE International Conference on Computer Vision (ICCV)*.

Ramanan, D. (2006). Learning to parse images of articulated bodies. In *Proceedings of Neural Information Processing Systems* (pp. 1129–1136). Cambridge: MIT Press.

Savarese, S., Winn, J., & Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision (ECCV)*.

Soille, P. (2003). *Morphological image analysis: Principles and applications* (2nd ed.). New York: Springer.

Souiai, M., Nieuwenhuis, C., Strekalovskiy, E., & Cremers, D. (2013). Convex optimization for scene understanding. In *ICCV Workshop on Graphical Models for Scene Understanding*.

Souiai, M., Strekalovskiy, E., Nieuwenhuis, C., & Cremers, D. (2013). A co-occurrence prior for continuous multi-label optimization. In *Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*.

Strekalovskiy, E., & Cremers, D. (2011). Generalized ordering constraints for multilabel optimization. In *IEEE International Conference on Computer Vision (ICCV)*.

Strekalovskiy, E., Goldluecke, B., & Cremers, D. (2011). Tight convex relaxations for vector-valued labeling problems. In *IEEE International Conference on Computer Vision (ICCV)*.

Strekalovskiy, E., Nieuwenhuis, C., & Cremers, D. (2012). Nonmetric priors for continuous multilabel optimization. In *European Conference on Computer Vision (ECCV)*.

Toeppe, E., Nieuwenhuis, C., & Cremers, D. (2013). Relative volume constraints for single view reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Toeppe, E., Oswald, M. R., Cremers, D., & Rother, C. (2010). Image-based 3d modeling via cheeger sets. In *Asian Conference on Computer Vision (ACCV)*.

Vezhnevets, A., Ferrari, V., & Buhmann, J. M. (2011). Weakly supervised semantic segmentation with a multi-image model. In *IEEE International Conference on Computer Vision (ICCV)*.

Wang, L., Shi, J., Song, G., & Shang, I. F. (2007). Object detection combining recognition and segmentation. In *Asian Conference on Computer Vision (ACCV)*.

Yao, J., Fidler, S., & Urtasun, R. (2012). Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zach, C., Gallup, D., Frahm, J. M., & Niethammer, M. (2008). Fast global labeling for real-time stereo using multiple plane sweeps. In *Vision, Modeling and Visualization Workshop (VMV)*.