Multi-cue mid-level grouping

Tom Lee, Sanja Fidler, Sven Dickinson

University of Toronto
{tshlee,fidler,sven}@cs.toronto.edu

Abstract. Region proposal methods provide richer object hypotheses than sliding windows with dramatically fewer proposals, yet they still number in the thousands. This large quantity of proposals typically results from a diversification step that propagates bottom-up ambiguity in the form of proposals to the next processing stage. In this paper, we take a complementary approach in which mid-level knowledge is used to resolve bottom-up ambiguity at an earlier stage to allow a further reduction in the number of proposals. We present a method for generating regions using the mid-level grouping cues of closure and symmetry. In doing so, we combine mid-level cues that are typically used only in isolation, and leverage them to produce fewer but higher quality proposals. We emphasize that our model is mid-level by learning it on a limited number of objects while applying it to different objects, thus demonstrating that it is transferable to other objects. In our quantitative evaluation, we 1) establish the usefulness of each grouping cue by demonstrating incremental improvement, and 2) demonstrate improvement on two leading region proposal methods with a limited budget of proposals.

1 Introduction

Casting object recognition as object detection diminishes the need for bottom-up grouping: a high-level model does not need the help of weaker mid-level and low-level cues to locate the object. However, as the level of ambiguity rises with the number of possible objects, the more prohibitive it becomes to exhaustively search over object detectors in a cluttered scene. This motivates the role of bottom-up cues for achieving a reduction in search complexity.

Bottom-up grouping has re-emerged in the form of class-independent *region proposals* [1, 2] which are increasingly combined with object detectors and have been shown to improve performance on competitive challenges [3]. Region proposal methods typically start with a generation stage that uses a bottom-up grouping algorithm to output a diverse set of proposals, which are then passed to a ranking stage where they are evaluated by a trained scoring function. The ranked proposals have richer structure than sliding windows, which are typically fixed in aspect ratio, and have higher precision than sliding windows, whose proposals number in the millions. In contrast, region proposal methods achieve state-of-the-art results with only thousands of proposals.

Region proposal methods forward bottom-up ambiguity from the generation stage to the ranking stage in the form of proposals, at which point stronger cues

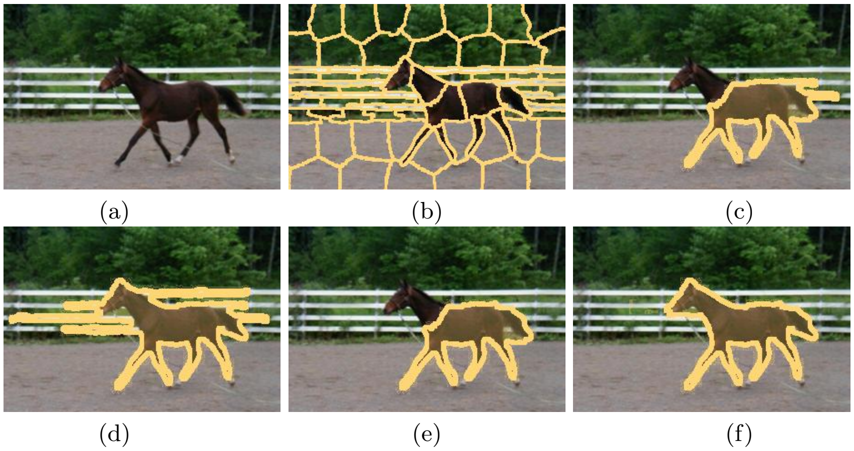

Fig. 1. Given an input image as shown in (a), our method first oversegments into superpixels in (b), which are to be grouped into regions based on a combination of perceptual grouping cues. In this example, both the horse and the fence are relatively homogeneous in color and exhibit contrasting boundaries, however the horse's neck is slightly darker than its torso. As shown in (c), low-level appearance alone oversegments the horse at the neck where a large gap in contour is attempted. When including contour closure in (d), the boundary correctly encloses the head, but elsewhere strays along the fence. Conversely in (e), including symmetry without closure separates the fence from the horse, but fails to enclose the head. With closure and symmetry together in (f), the entire horse is correctly segmented.

are available to reduce the ambiguity. Unlike hierarchy-based models [4], proposals are often explicitly isolated from object class labels. Typical methods like [1, 2], however, rely on only low-level appearance and contour cues to generate proposals, and as a result must diversify their proposals in large quantities to preserve recall. In this paper, we present a complementary approach to diversification that uses mid-level grouping cues to resolve ambiguity at an early stage to avoid the need to generate proposals in excessive quantities.

By approaching the problem as figure-ground separation, we draw on a large body of work in perceptual grouping. Mid-level cues capture non-accidental relations between image elements that are exhibited by all objects. They are less specific than a high-level object model, yet more discriminative than low-level cues like appearance similarity and contour continuity. Here we highlight two mid-level cues of interest:

Closure [5, 6] is a regularity that favors regions that are enclosed by strong contour evidence along the boundary. Bottom-up approaches to finding closure vary in the types of cues used, and may include continuity and convexity. The problem is often cast as finding a cycle of graph edges in a very large space, and is exacerbated when allowing for gaps in the closure (an illustrative example being the Kanizsa triangle).

Symmetry [7–9] is a ubiquitous and powerful regularity with scope that spans entire objects or their parts. Since the early days, perceptual grouping research has produced such varied representations as the medial axis transform [10], generalized cylinders [11], superquadrics [12], and geons [13]. Later approaches applied symmetry toward cluttered and occluded image domains, which present the challenge of searching for symmetrically related elements in an intractably large space.

Like other bottom-up cues, closure and symmetry govern the perception of figure and ground. Our method, as illustrated in Figure 1, groups regions by leveraging mid-level and low-level cues in combination. An input image (a) is oversegmented into superpixels (b) to be grouped together into regions. The example shown contains a horse as foreground, for which multiple grouping cues will help to separate from the background. Relying on a limited number of cues, as subsequently shown, may result in a segmentation that is overly sensitive to detailed changes in the image. In (c), low-level appearance alone oversegments the horse at the slightly darker neck, while jumping a large gap in contour. Including contour closure in (d) attracts the boundary to pixels with strong contour evidence and encloses the head, but elsewhere strays along the fence. Symmetry is a regularity that groups objects, such as the fence, into its coherent parts, but as shown in (e), does not group the head with the horse. In (f), closure and symmetry combine their strengths to correctly segment the horse.

Mid-level cues extend beyond any particular object, and symmetry and closure, in particular, are ubiquitous over all objects. Since our model is aware of objects only at the mid-level and unaware of their specific appearance, the model can easily transfer from one object to another. In this paper, as a case in point, we learn our model on the Weizmann Horse Dataset (WHD) [14], and then apply it to diverse non-horse objects from the Weizmann Segmentation Dataset (WSD) [15]. Quantitative experiments are performed on WHD to 1) establish the usefulness of each cue by demonstrating improvement as they are incrementally added, and to 2) demonstrate improvement on two leading region proposal methods with a limited budget of proposals. The contributions of our paper are summarized as follows:

1. **Perceptual search.** We focus on the front-end stage where the generation of region proposals is driven by bottom-up grouping. We argue that stronger mid-level cues play an important role in reducing the number of proposals.
2. **Mid-level cue combination.** We improve upon previous approaches that lack mid-level knowledge or combine only one mid-level cue with low-level cues, by leveraging the combination of mid-level closure and mid-level symmetry to group regions together.
3. **Trained cue combination:** While perceptual grouping methods often make ad hoc grouping decisions, we capture all cues in a single energy function and jointly learn their weighted combination.

2 Related work

Viewing region proposals as object hypotheses for recognition, we begin by broadening our scope to include methods designed for sliding window detectors.

Among these, the *objectness* detector of Alexe *et al.* [16] computes low-level features on superpixels [17] to score sampled image boxes. *Selective search* of Uijlings *et al.* [2] outputs boxes that bound regions generated from agglomerative clustering of superpixels [17]. The method accumulates a pool of regions over each step of region-merging until all regions are merged together, and ensures diversity by pooling results over multiple color and texture feature spaces. The method is very fast, yet is based on low-level appearance alone.

Arbelaez *et al.* [18] produces regions by merging superpixels of [19] over multiple scales. The method considers a limited number of all pairs, triples, and quadruples of adjacent superpixels. Our approach is different in that we operate on a single layer of compact superpixels, and define a set of low-level and mid-level cues that quantify the likelihood of grouping.

The *shape sharing* method of Kim & Grauman [20] matches part-level regions in a given image to a bank of exemplars, which project object-level information back into the image to help with segmentation. The *category-independent proposals* of Endres & Hoiem [21] develops a CRF model to label superpixels based on segment seeds. The resulting region proposals are ranked using structured learning on grouping cues. The energy potentials are pairwise and submodular, and inference is done by graph cuts. While we use a similar procedure to generate regions, we combine mid-level cues at the front-end without seeding from a fixed hierarchical segmentation.

The *CPMC* method of Carreira & Sminchisescu [1] generates regions directly from the image rather than deriving them from a fixed segmentation. The method solves multiple parametric min-cut instances over color seeds. Regions are re-ranked by regressing on overlap with region-scoped features, including mid-level features such as convexity and eccentricity. The emphasis is on ranking rather than the front-end grouping, which samples color seed models over millions of pixels. Our approach is qualitatively different from the above methods as we focus on bottom-up grouping, however our mid-level front-end is complementary to the ranking stage.

Viewing region proposals as figure-ground labeling calls on a large literature covering low-level and mid-level Gestalt cues. Rather than covering methods on individual mid-level cues like symmetry [7–9] and closure [5, 6], we consider holistic approaches that combine low- and/or mid-level cues. The *region competition* approach of Zhu & Yuille [22] combines the objectives of snakes and region growing into a single Bayes criterion, effectively integrating the relative strengths of contour-based and region-based cues. An algorithm for optimizing the new criterion was introduced, however only guaranteed convergence to a local minimum. Our approach differs in using superpixels which, providing access to both contours and regions, serves as a convenient basis for combining their respective cues, independently from the optimization approach.

Cue combination is alternately formulated as a linear combination of terms that make up a cost or scoring function. Graph-based image partitioning [17, 23] requires an affinity function to be specified between pairs of pixels and therefore falls under this category. For example, the *intervening contour* method of

Leung & Malik [24] includes a contour-based term into the appearance-based affinity and solves the normalized cut problem. Like [24], we combine cues in a linear combination of terms, but differ in the overall grouping approach and use different cues on superpixels.

Inspired by random field models, the *cue integration* method of Ren *et al.* [25] develops an energy function that integrates appearance similarity, contour continuity, contour closure, and object familiarity on triangular tokens. The model was trained and solved using loopy belief propagation. Like [25], we combine multiple grouping cues over adjacent regions, but we take the approach of expressing the energy potentials in a form that allows efficient and exact solutions.

Our approach is most similar to Levinshtein *et al.* [26], which elegantly formulated contour closure as finding minimum energy labelings, and used parametric min-cut to find globally optimal solutions. A gap cost was trained on superpixel boundary features and incorporated into a gap-to-area ratio cost. We differ from [26] by combining multiple cues, among which contour closure counts as only one, and furthermore we learn to combine cues in a random field energy model.

3 Approach overview

We develop an energy function over superpixel labelings that captures a combination of low-level and mid-level grouping cues. In Section 4, we motivate the cues of low-level appearance, mid-level closure, and mid-level symmetry from perceptual grouping principles and define their corresponding energy potentials. We use a mathematical form that is flexible enough to accommodate additional cues, yet conforms to a structure that can be exploited to obtain efficient and exact solutions. In Section 5, we introduce a scaling term in the energy that represents ambiguity in scale, and use it to obtain multiple solutions. Section 6 formulates the loss-based framework with which we train the weights of the energy function. We present and discuss results in Section 7 and conclude in Section 8.

4 Grouping cues

Our method operates on superpixels as grouping primitives from which regions are composed. Superpixels provide a rich topology of regions and boundaries on which a diverse set of cues can be defined to capture different grouping relations. Specifically, an input image $\mathbf{x}$ is oversegmented into P superpixels, where each superpixel p is assigned a binary label $y_p \in \{1 = \text{figure}, 0 = \text{ground}\}$. The labeling space $\mathcal{Y} = \{1,0\}^P$ contains all possible vectors $\mathbf{y} = \{y_1, \ldots, y_P\}$ of superpixel labels and thus represents all possible groupings. An energy function $E(\mathbf{y}; \mathbf{x})$ is defined on $\mathcal{Y}$ that favors labelings based on a combination of cues observed on the image $\mathbf{x}$, and captures this combination as a decomposition into potentials corresponding to different cues:

$$E(\mathbf{y}; \mathbf{x}) = \sum_{cue} \sum_{I \in \mathcal{N}^{cue}} E_I^{cue}(\mathbf{y}_I; \mathbf{x}_I) \tag{1}$$

In (1), *cue* varies over low-level appearance (*app*), mid-level closure (*clo*), and mid-level symmetry (*sym*). The set $\mathcal{N}^{cue}$ of neighborhoods for a particular cue defines the local subsets of superpixels on which the cue is repeatedly observed. Potentials in our model are restricted to pairwise order. By finding a labeling that globally minimizes the energy, we obtain a region that exhibits strong grouping relations. In this section, we discuss the contributions of the cues of symmetry, closure, and appearance and define their corresponding energy potentials.

4.1 Appearance similarity

Similarity is a basic perceptual grouping cue that we capture in the form of color and texture similarity. We note that even objects of heterogeneous appearance are often composed of homogeneous parts. For each superpixel p, we compute a d-dimensional normalized histogram descriptor $\mathbf{h}^p$ that summarizes its appearance. We then compute the similarity between a pair p, q of adjacent superpixels using the histogram intersection kernel:

$$s^{pq} = \sum_{i=1}^{d} \min(h_i^p, h_i^q).$$

Color and texture are captured with different histograms $\mathbf{h}_c, \mathbf{h}_t$ which are computed in the manner of Uijlings *et al.* [2] using multiple color channels and SIFT-like features. Similarity is computed for both histograms to obtain the two-dimensional feature:

$$\phi_{pq}^{app}(\mathbf{x}) = (s_c^{pq}, s_t^{pq}).$$

The pairwise appearance potential for each adjacent pair p, q combines the cues and is defined as follows:

$$E_{pq}^{app}(\mathbf{y}_{pq}; \mathbf{x}) = \begin{cases} \mathbf{w}_{app}^T \phi_{pq}^{app}(\mathbf{x}) & y_p \neq y_q \\ 0 & y_p = y_q \end{cases} \tag{2}$$

4.2 Contour closure

Contour closure is a key challenge of perceptual grouping. One of the key ingredients of closure is strong contour evidence along the boundary that separates figure from ground. Since we prefer boundaries that avoid large gaps of contour (weak evidence), we define for any given labeling $\mathbf{y}$ the gap cost $G(\mathbf{y})$ in terms of the corresponding region's boundary $\partial(\mathbf{y})$:

$$G(\mathbf{y}; \mathbf{x}) = \sum_{x \in \partial(\mathbf{y})} g(x).$$

This cost accumulates a positive gap $g(x)$ over all boundary pixels $x \in \partial(\mathbf{y})$. We compute $g(x) \in [0, 1]$ at every boundary pixel using the trained measure of [26], which accounts for discrepancy between contour map and superpixel boundaries in location and orientation.

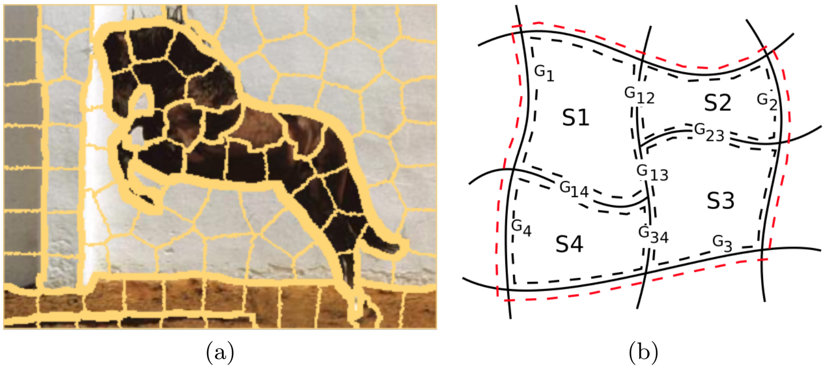

Fig. 2. To support the cue of mid-level closure, contour evidence is computed along superpixel boundaries as shown in (a), where thickness indicates the degree of contour evidence (lack of gap) [26]. In (b), the gap cost $G(\mathbf{y})$ for a hypothetical labeling $\mathbf{y}$ over the corresponding region's boundary $\partial(\mathbf{y})$ is shown in dashed red and consists of superpixels S1-S4. Unary potentials sum gap along the corresponding boundaries G1-G4, and pairwise potentials sum gap along the shared boundaries G12-G34. The total gap $G(\mathbf{y})$ along the dashed red is obtained by subtracting twice the pairwise potentials from the unary potentials. (We thank the authors of [26] for permission to reproduce figure (b)).

We directly incorporate $G(\mathbf{y})$ into our energy function by expressing it in terms of unary and pairwise potentials over $\mathbf{y}$. We encode the potentials as in [26], for which a schematic example is provided in Figure 2. Unary potentials are defined to sum gap along the corresponding superpixel's boundary $\partial(p)$ when $y_p = 1$. Pairwise potentials between p and q sum gap only along the boundary $\bar{\partial}(p, q)$ shared by *both* superpixels, when $y_p = y_q = 1$:

$$E_p^{clo}(y_p) = \begin{cases} \sum_{x \in \partial(p)} g(x) & y_p = 1 \\ 0 & y_p = 0 \end{cases} \quad E_{pq}^{clo}(\mathbf{y}_{pq}) = \begin{cases} \sum_{x \in \bar{\partial}(p,q)} g(x) & y_p = y_q = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

As illustrated in Figure 2(b), unary potentials sum gap along their superpixel boundaries. For a region consisting of a single superpixel, the unary potential reflects the correct gap cost. However, for a region consisting of multiple superpixels, simply summing the corresponding unary potentials will double count the gaps along the boundaries shared by adjacent superpixels in the region, which are exactly those counted by the pairwise potentials. The gap $G(\mathbf{y})$ along the true boundary of the region can thus be easily expressed as the sum of the unary potentials, minus twice the pairwise potentials:

$$E^{clo}(\mathbf{y}; \mathbf{x}) = w_{clo} \cdot \left(\sum_p E_p^{clo}(y_p; \mathbf{x}) - 2 \sum_{p,q} E_{pq}^{clo}(\mathbf{y}_{pq}; \mathbf{x}) \right) \tag{4}$$

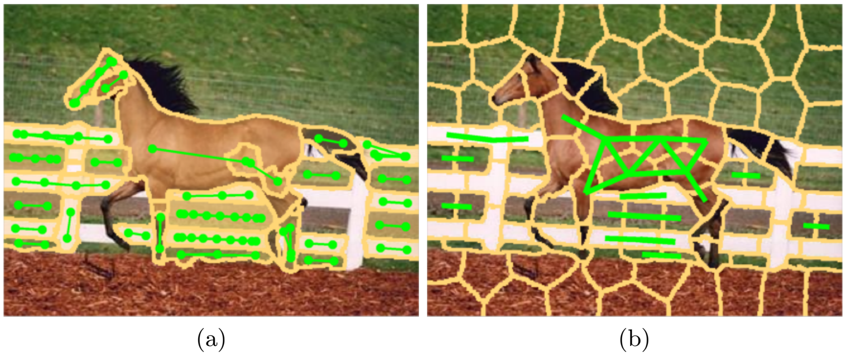

Fig. 3. Symmetric parts detected by [27] as sequences of medial points represented as region masks, as shown in (a). In (b), straight lines indicate strong pairwise affinities between superpixels that belong to the same symmetric part.

4.3 Symmetry

Symmetry relates together local features that span the entire object or its parts and, as such, is a powerful mid-level cue. Its large spatial scope, however, makes the associated grouping problem combinatorially hard. In the context of our representation in the labeling space $\mathcal{Y}$, the region corresponding to an object or its part can be composed from any number of superpixels, and thus induces dependencies of arbitrarily high order.

Our method draws on the approach of Lee *et al.* [27] for finding symmetrically related features, which circumvents the above difficulty by leveraging the scope of large superpixels. By operating on successively coarser superpixels, pairwise combinations are able to cover successively larger regions, effectively achieving higher orders of dependency. This allows local sections of symmetry of the same object part to be composed from a *sequence* of pairwise superpixels at the correct scale. Furthermore, [27] finds optimal sequences of superpixels that lie along the symmetry axes of object parts, as shown in Figure 3.

We incorporate the symmetry cue in the above form into our method by favoring the grouping of superpixels that belong, with high likelihood, to the same symmetric part. In practice, we run the sequence optimization of [27] independently on multiscale superpixels to obtain a set S of symmetric parts, as shown in Figure 3(a), and define pairwise potentials that favor grouping of superpixels that belong to the same symmetric part, as shown in 3(b).

For each pair of adjacent superpixels p, q, we define the feature:

$$\phi_{pq}^{sym}(\mathbf{x}) = \max_{s \in S(p,q)} \text{score}(s),$$

which takes on the score of the best scoring symmetric part $s \in S(p,q)$, where $S(p,q) \subseteq S$ is the subset of symmetric parts for which the overlap with p and q both exceed $\tau = 0.75$. When $S(p,q)$ is empty, the feature takes on a value of zero. The value score$(s) \in [0,1]$ is the part's detection score, which we interpret as positive grouping evidence. We perform non-maximum suppression over all superpixels pairs so that each pairwise relation is influenced by at most one symmetric part. The symmetry potential is defined for each pair (p,q) of adjacent superpixels as:

$$E_{pq}^{sym}(\mathbf{y}_{pq};\mathbf{x}) = \begin{cases} \mathbf{w}_{sym}^T \phi_{pq}^{sym}(\mathbf{x}) & y_p \neq y_q \\ 0 & y_p = y_q. \end{cases} \tag{5}$$

5 Figure-ground labeling

We incorporate the potentials corresponding to the grouping cues into our final energy function as follows:

$$E(\mathbf{y}) = \sum_{p,q} E_{pq}^{app}(\mathbf{y}) + \sum_{p} E_{p}^{clo}(\mathbf{y}) - 2\sum_{p,q} E_{pq}^{clo}(\mathbf{y}) + \sum_{p,q} E_{pq}^{sym}(\mathbf{y}) + \lambda \sum_{p} \phi_p(\mathbf{y}). \tag{6}$$

In (6), the grouping cues are rescaled by a scaling potential $\phi_p(\mathbf{y})$ by a factor of $\lambda > 0$ that is defined as follows:

$$\phi_p(\mathbf{y}) = \begin{cases} -\text{area}(p) & y_p = 1 \\ 0 & y_p = 0. \end{cases} \tag{7}$$

The scaling potential removes trivial solutions associated with the empty grouping with zero energy. Furthermore, as λ increases, the scaling potential favors labelings of larger area, and thus λ adjusts the energy's preference for regions of smaller or larger scale.

To minimize (6), we rewrite it as a sum of unary and pairwise potentials:

$$E(\mathbf{y};\mathbf{x}) = \sum_{p} \mathbf{w}_1^T \phi_p^{\lambda}(\mathbf{y},\mathbf{x}) + \sum_{p,q} \mathbf{w}_2^T \phi_{pq}(\mathbf{y},\mathbf{x}), \tag{8}$$

noting that the pairwise potentials are submodular when weights are non-negative (features are non-negative). When λ is fixed, (8) can be minimized efficiently with a maxflow algorithm. In our model, λ is an unknown variable that represents the scale of an object, and so we minimize (8) for all values $\lambda \in \Lambda$, for $\Lambda \subset \mathbb{R}$. This is known as the parametric maxflow problem [28], which can be shown to yield a finite number of solutions as λ varies over Λ. The set of globally optimal solutions can be found with a linear number of calls to the maxflow algorithm. We use $\Lambda = [0,1]$ to yield a dozen solutions on average per image, thereby obtaining multiple proposals varying in scale.

6 Learning

We train the weights of the energy function (8) by incorporating it into the Structured SVM framework. The framework is instantiated with the loss function:

$$\Delta(\hat{\mathbf{y}},\mathbf{y}) = \frac{1}{\text{area}(\mathbf{x})} \sum_{p} \text{area}(p) \cdot \phi_p^{\Delta}(\hat{y}_p, y_p) \qquad \phi_p^{\Delta}(\hat{y}, y) = \begin{cases} 1-\alpha_p & \hat{y} = 1 \\ \alpha_p & \hat{y} = 0 \end{cases} \tag{9}$$

where $\alpha_p \in [0,1]$ is the fraction of pixels inside superpixel p labeled by the ground truth pixel mask. Weights are optimized using StructSVMCP [29] and constrained to be non-negative. We note that the learning step assumes that the loss for a particular example is obtained by minimizing the corresponding energy with a particular value of λ. For simplicity, we have fixed $\lambda = 0.01$ for all training examples. During testing, however, we vary λ over $\mathit{\Lambda}$ for each example.

7 Evaluation

A key point of our approach is that our model being mid-level enables it to directly transfer from one object class to another. To illustrate this point, we use the Weizmann Horse Dataset (WHD) [14] to build our model, while applying it on diverse non-horse objects from the Weizmann Segmentation Dataset (WSD) [15]. Section 7.2 describes the qualitative results obtained on WSD. We additionally perform quantitative experiments to study the individual contributions of our grouping cues, and to demonstrate an improvement over two leading region proposal methods. Results are presented in Sections 7.1 and 7.3, respectively.

Contained in WHD are 328 images, each annotated with a ground truth mask. We train on the first 200 images, and hold out the remainder for test. As an evaluation metric, we compute the average best overlap [2]:

$$\mathcal{O}(\mathcal{G}, \mathcal{R}; k) = \frac{1}{|\mathcal{G}|} \sum_{(g,i) \in \mathcal{G}} \max_{r \in \mathcal{R}(i;k)} o(r; g),$$

where $\mathcal{G}$ and $\mathcal{R}$ are the ground truth and region masks, respectively, and the quantity k is the number of top-ranked proposals. Intersection-over-union overlap between a region r and the ground truth mask g is denoted by $o(r; g)$. We plot overlap against k to measure the trade-off between overlap and k.

7.1 Cue combination

We study the effect of incrementally combining the cues of appearance, closure and symmetry, by including their respective potentials in the energy function (6). Each cue observes a different type of grouping evidence, and we expect the best result from combining the strengths of all cues. Figure 4 shows the effect of incrementally adding closure and symmetry to appearance, as well as using mid-level cues without appearance. We observe that closure and appearance work well together, while symmetry helps for all combinations. The results confirm our hypothesis that each cue individually contributes useful information, with the best result from combining all cues. Our symmetry cue contributed a smaller than expected improvement on WHD. We expect symmetry's contribution to be better reflected in more challenging datasets of objects whose regions cannot be as easily computed with the remaining cues alone.

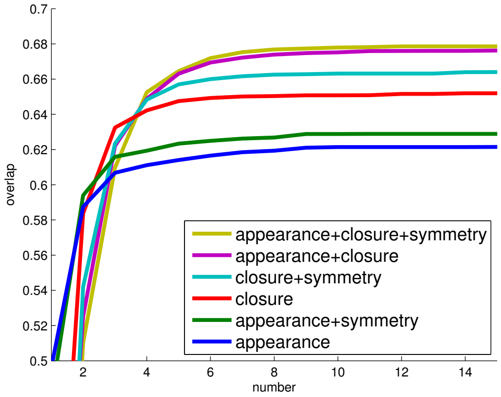

Fig. 4. Improvement in recall as grouping cues are incrementally added to the energy.

7.2 Qualitative results

We present qualitative results for a diverse set of objects in Figure 6, where each row shows the top proposed regions produced from a given input image, along with the corresponding ground truth mask. Our method successfully separates horses from cluttered and occluded backgrounds. We observe that alternative regions often arise when there are spurious contours, particularly within the horse and shadows under the horse. False negative contours, however, can cause undersegmentation, *e.g.*, in row 4. Symmetry of occluded fences is often sufficient to prevent undersegmentation. We note that while our appearance cue favors grouping regions of similar color, it does not penalize regions of heterogeneous color and correctly segments the horse in row 5. The remaining rows show results on different objects from WSD and demonstrate that our method is class-independent, and that our mid-level cues trained on WHD transfers well to objects of different classes.

7.3 Comparison with region proposals

We demonstrate the advantage of our mid-level method with respect to Selective Search [2] and CPMC [1] in Figure 5. For comparison with [2], we have measured overlap with respect to the agglomerated regions (rather than their bounding boxes), pooled over color types, similarity measures, and the parameter of [17]. The quantitative comparison demonstrates an improvement on [2] with a budget of a hundred proposals. We note that our method focuses on resolving ambiguity and generates 20-30 proposals per image. In contrast, [2] relies on diversity of proposals and requires over 100 proposals to achieve the same recall. Our method is thus more effective for a limited budget of proposals.

For comparison with [1], we have measured overlap with their regions produced using color seeds, where a color seed model is fit to sampled locations.

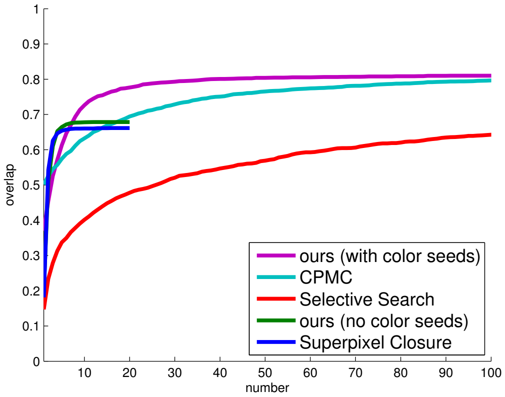

Fig. 5. Improvements over CPMC [1] and Selective Search [2] with a limited budget of proposals, and improvement over Superpixel Closure [26]. Our method is evaluated with and without color seeds. See text for details.

For this comparison, we have also added color seeds to our energy function (6). Specifically, for any given test image, we fit a Gaussian mixture model to the image's RGB distribution to obtain a compact set of color seed models corresponding to each mixture component (we obtain 4-6 clusters per image). This differs from [1] which densely samples color seeds over a grid. For each pair of color seed as a foreground-background hypothesis, we bias our energy function (6) with a unary potential that scores the corresponding superpixel's log likelihood ratio between the foreground seed and the background seed, as done in [1]. Parametric min-cut is solved for each pair of color seed, and the resulting regions are pooled with the original (unbiased) regions, obtaining several hundred proposals per image. Our method with color seeds improves on [1] with a budget of a hundred proposals.

8 Conclusion

Bottom-up grouping is regaining momentum as a counterpart to object detection, and is a promising area in which to explore the importance of mid-level grouping cues. Mid-level cues are ubiquitous and transcend individual object classes, yet can be leveraged effectively only in combination. We have presented a method to combine appearance, closure, and symmetry, and demonstrated the usefulness of each cue. We have also demonstrated the effectiveness of using mid-level cues to resolve ambiguity with a limited budget of proposals, and shown that our model complements diversification techniques when a large number of proposals is affordable.

References

1. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric min-cuts. PAMI **34** (2012) 1312–1328
2. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV **104** (2013) 154–171
3. Fidler, S., Mottaghi, R., Yuille, A., Urtasun, R.: Bottom-up segmentation for top-down detection. CVPR (2013) 3294–3301
4. Fidler, S., Boben, M., Leonardis, A.: Learning a hierarchical compositional shape vocabulary for multi-class object representation. ArXiv:1408.5516 (2014)
5. Elder, J., Zucker, S.: Computing contour closure. ECCV (1996) 399–412
6. Jacobs, D.: Robust and efficient detection of convex groups. PAMI (1996)
7. Loy, G., Eklundh, J.: Detecting symmetry and symmetric constellations of features. ECCV (2006) 508–521
8. Mohan, R., Nevatia, R.: Perceptual organization for scene segmentation and description. PAMI **14** (1992) 616–635
9. Tsogkas, S., Kokkinos, I.: Learning-based symmetry detection in natural images. ECCV (2012)
10. Blum, H.: A transformation for extracting new descriptors of shape. Models for the perception of speech and visual form **19** (1967) 362–380
11. Binford, T.: Visual perception by computer. ICSC (1971)
12. Pentland, A.: Perceptual organization and the representation of natural form. AI (1986)
13. Biederman, I.: Human image understanding: Recent research and a theory. CVGIP (1985)
14. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. ECCV (2002)
15. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. CVPR (2007)
16. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? CVPR (2010) 73–80
17. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV **59** (2004) 167–181
18. Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. CVPR (2014)
19. Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. PAMI **33** (2011) 898 – 916
20. Kim, J., Grauman, K.: Boundary preserving dense local regions. CVPR (2011)
21. Endres, I., Hoiem, D.: Category independent object proposals. ECCV (2010)
22. Zhu, S., Yuille, A.: Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. PAMI **18** (1996) 884–900
23. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI (2000)
24. Leung, T., Malik, J.: Contour continuity in region based image segmentation. ECCV (1998) 544–559
25. Ren, X., Fowlkes, C., Malik, J.: Cue integration for figure/ground labeling. NIPS (2005)
26. Levinshtein, A., Sminchisescu, C., Dickinson, S.J.: Optimal image and video closure by superpixel grouping. IJCV (2012)
27. Lee, T., Fidler, S., Dickinson, S.: Detecting curved symmetric parts using a deformable disc model. ICCV (2013)
28. Kolmogorov, V., Boykov, Y., Rother, C.: Applications of parametric maxflow in computer vision. ICCV **8** (2007)
29. Schwing, A., Fidler, S., Pollefeys, M., Urtasun, R.: Box in the box: Joint 3d layout and object reasoning from single images. ICCV (2013)

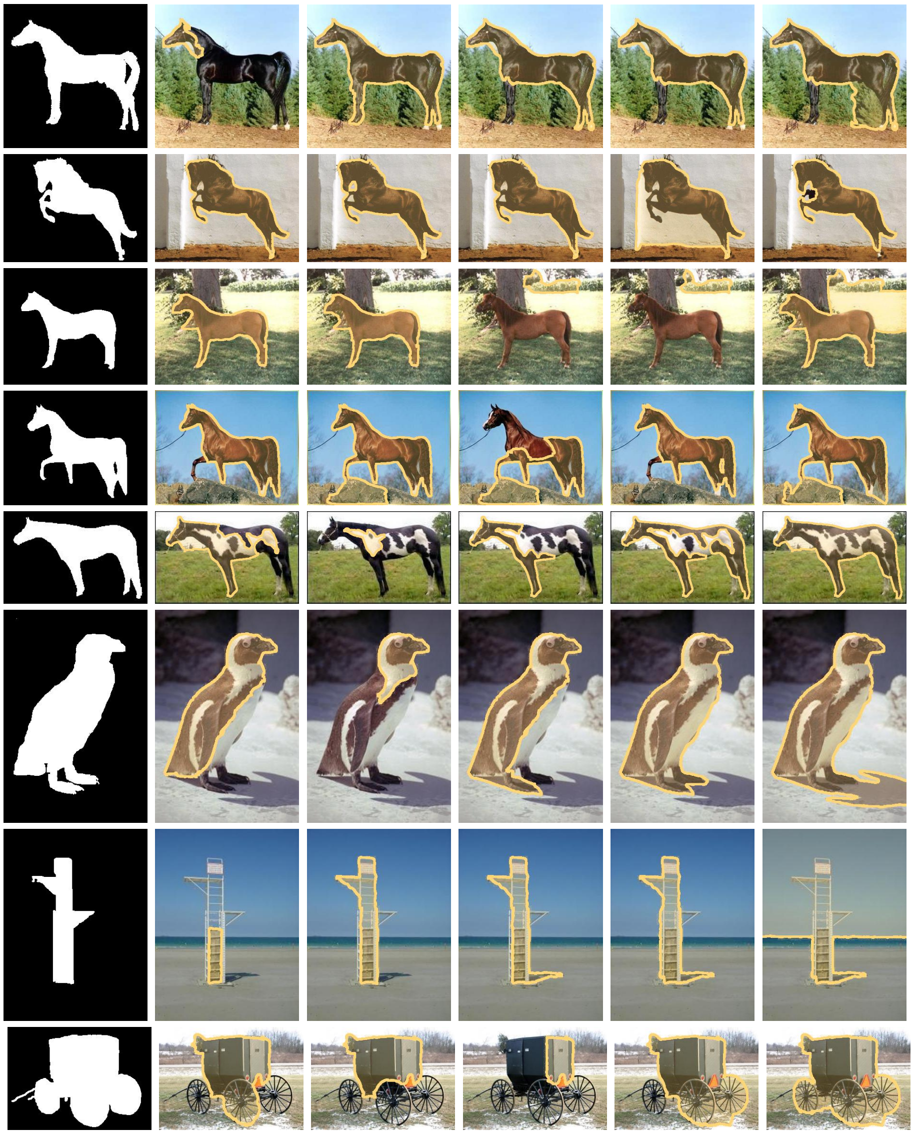

Fig. 6. Top region proposals from our method from different images. Leftmost column shows corresponding ground truth masks and remaining columns show region proposals. Rows 1-5 correspond to images from the Weizmann Horse Database (WHD), and rows 6-8 correspond to images from the Weizmann Segmentation Database (WSD). See text for details.