Affinity CNN: Learning Pixel-Centric Pairwise Relations for Figure/Ground Embedding

Michael Maire[1], Takuya Narihira[2], Stella X. Yu[3]
[1]TTI Chicago [2]Sony Corp. [3]UC Berkeley / ICSI

We build a system for simultaneous image segmentation and figure/ground organization by directly connecting a convolutional neural network (CNN) to a spectral embedding algorithm which produces a globally consistent scene interpretation. Training the CNN with a target appropriate for this inference procedure eliminates the need for the hand-designed intermediate stages, such as edge detection, typical of perceptual organization pipelines.

Angular Embedding (AE) [9] serves as our globalization framework. Previous work [5] establishes AE as a basis for joint segmentation and figure/ground organization. We follow in spirit, but employ major changes in order to achieve results of the high quality shown in Figure 1:

- We reformulate segmentation and figure/ground layering in terms of an energy model with pairwise forces between pixels. Pixels either bind together (group) or differentially repel (layer separation), with strength of interaction modulated by confidence in the prediction.
- We learn a CNN that predicts these interactions across multiple distance scales and use an efficient solver [6] for spectral embedding.

Figure 2 details our architecture for connecting a CNN with AE.

To arrive at this design, we first abstract the figure/ground problem to that of assigning each pixel p a rank $\theta(p)$, such that $\theta(\cdot)$ orders pixels by occlusion layer. Now assume we can obtain estimates of the relative order $\Theta(p,q)$ between many pairs of pixels p and q. The task is then to find $\theta(\cdot)$ that agrees as best as possible with these pairwise estimates. AE addresses this optimization problem by minimizing error:

$$\varepsilon = \sum_p \frac{\sum_q C(p,q)}{\sum_{p,q} C(p,q)} \cdot |z(p) - \tilde{z}(p)|^2 \tag{1}$$

where $C(p,q)$ accounts for possibly differing confidences in the pairwise estimates and $\theta(p)$ is replaced by $z(p) = e^{i\theta(p)}$. As Figure 3 shows, this permits interpretation of $z(\cdot)$ as an embedding into the complex plane, with desired ordering $\theta(\cdot)$ corresponding to absolute angle. $\tilde{z}(p)$ is defined as the consensus embedding location for p according to its neighbors and Θ:

$$\tilde{z}(p) = \sum_q \tilde{C}(p,q) \cdot e^{i\Theta(p,q)} \cdot z(q) \tag{2}$$

$$\tilde{C}(p,q) = \frac{C(p,q)}{\sum_q C(p,q)} \tag{3}$$

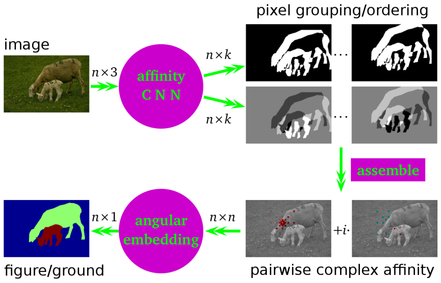

Figure 2: **Architecture.** A trained CNN predicts grouping and ordering relations between each of the n image pixels and its neighbors at k displacements across a fixed stencil pattern. We assemble these $n\times2k$ pixel-centric relations into a sparse $n\times n$ complex affinity matrix. We feed the pairwise affinity matrix into Angular Embedding [9] for global integration, producing an eigenvector representation that reveals segmentation and figure-ground organization: we know not only which pixels go together, but also which pixels go in front.

Relaxing the unit norm constraint on $z(\cdot)$ yields a generalized eigenproblem:

$$Wz = \lambda Dz \tag{4}$$

with $D = \mathrm{Diag}(C1_n)$ and $W = C \bullet e^{i\Theta}$, where $\bullet$ denotes Hadamard product.

For Θ everywhere zero ($W = C$), this eigenproblem is identical to the spectral relaxation of Normalized Cuts [8], in which the second and higher eigenvectors encode grouping [1, 8]. With nonzero entries in Θ, the first of the now complex-valued eigenvectors is nontrivial and its angle encodes rank ordering while the subsequent eigenvectors still encode grouping [5]. We use the same decoding procedure as [5] to read off this information. We also recover boundaries and segmentation from the embedding by taking the (spatial) gradient of eigenvectors and applying the watershed transform.

It remains to define the pairwise pixel relationships $C(p,q)$ and $\Theta(p,q)$. Figure 4 illustrates possible transitions between p and q. Selecting the most likely, the probabilities of erroneously binding p and q into the same region, transitioning to figure, or transitioning to ground are:

$$E_B(p,q) = b(p,q) \tag{5}$$

$$E_F(p,q) = 1 - (1-e(p))b(p,q)(1-e(q))f(p,q) \tag{6}$$

$$E_G(p,q) = 1 - (1-e(p))b(p,q)(1-e(q))g(p,q) \tag{7}$$

where $e(p)$ and $b(p,q)$ denote the probability of an edge at p and a boundary somewhere between p and q, respectively. $f(p,q)$ and $g(p,q)$ are conditional probabilities of relative figure and ground, given that p and q are in separate regions. Note $g(p,q) = 1 - f(p,q)$. Figure/ground repulsion forces act long-range and across boundaries. We convert to confidence via exponential reweighting (σ_b and $\sigma_f = \sigma_g$ control scaling), and apply a rotational action by fixed angle ϕ for figure/ground transitions, obtaining affinities:

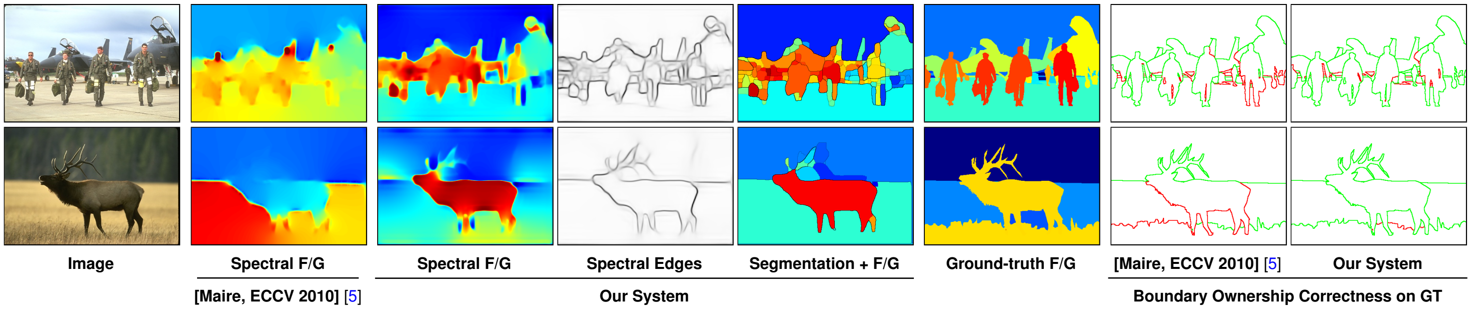

Figure 1: **Segmentation and figure/ground results.** ***Left:*** While both utilize spectral embedding, our improved energy model and CNN-based predictors yield significant performance gains over the prior work of Maire [5]. Spectral F/G shows predicted per-pixel figure/ground ordering. Compare the strong lower-region bias for figure of [5] to our correct extraction of foreground objects. Spectral edges show soft boundary strength encoded by the embedding. These boundaries generate a hierarchical segmentation [1], one level of which we display with per-pixel figure/ground averaged over regions. ***Right:*** Averaging instead over ground-truth regions, we can project a predicted figure/ground ordering onto the ground-truth segmentation. For boundaries separating regions with different ground-truth figure/ground layer assignments, we check whether the predicted owner (more figural region) matches the owner according to the ground-truth. The rightmost two columns mark correct boundary ownership predictions in green and errors in red.

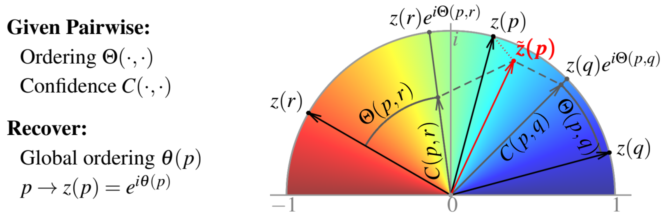

Figure 3: **Angular Embedding** [9]**.** Given (C,Θ) capturing pairwise relationships between nodes, the Angular Embedding task is to map those nodes onto the unit semicircle, such that their resulting absolute positions respect confidence-weighted relative pairwise ordering (Equation 1). Relative ordering is identified with rotation in the complex plane. For node p, $\theta(p) = \arg(z(p))$ recovers its global rank order from its embedding $z(p)$.

$$W_B(p,q) = \exp(-E_B(p,q)/\sigma_b) \tag{8}$$

$$W_F(p,q) = \exp(-E_F(p,q)/\sigma_f)\exp(i\phi) \tag{9}$$

$$W_G(p,q) = \exp(-E_G(p,q)/\sigma_g)\exp(-i\phi) \tag{10}$$

Combining these base cases into a single energy model yields affinity:

$$W(p,q) = W_B(p,q) + W_F(p,q) + W_G(p,q) \tag{11}$$

One can regard $W(p,q)$ as a sum of binding, figure transition, and ground transition forces acting between p and q. Learning to predict $e(p)$, $b(p,q)$, and $f(p,q)$ suffices to determine all components of W. For computational efficiency, we predict pairwise relationships between each pixel and its immediate neighbors across multiple spatial scales. As an adjustment prior to feeding W to the Angular Embedding solver of [6], we enforce Hermitian symmetry by assigning: $W \leftarrow (W + W^*)/2$.

Supervised training of our system proceeds from a collection of images and associated ground-truth, $\{(I_0,S_0,R_0),(I_1,S_1,R_1),\ldots\}$. Here, I_k is an image defined on domain $\Omega_k \subset \mathbb{N}^2$. $S_k : \Omega_k \to \mathbb{N}$ is a segmentation mapping each pixel to a region id, and $R_k : \Omega_k \to \mathbb{R}$ is an rank ordering of pixels according to figure/ground layering. This data defines pairwise relationships:

$$\tilde{b}_k(p,q) = 1 - \delta(S(p) - S(q)) \tag{12}$$

$$\tilde{f}_k(p,q) = (sign(R(q) - R(p)) + 1)/2 \tag{13}$$

As $f(p,q)$ is a conditional probability, we only generate training examples $\tilde{f}_k(p,q)$ for pairs (p,q) satisfying $\tilde{b}_k(p,q) = 1$. We compute $e(\cdot)$ from $b(\cdot,\cdot)$. Figure 5 illustrates derivation of training signals from the annotation available on the Berkeley segmentation dataset (BSDS) [7].

Choosing a CNN to implement these predictors, we regard the problem as mapping an input image to a 48 channel output over the same domain. The 48 channels are predictors for $b(\cdot,\cdot)$ and $f(\cdot,\cdot)$ at each of 24 offsets (8 immediate neighbors across 3 scales). We use an AlexNet [4]-inspired design, augmented to include both coarse and fine receptive fields [2], and train with log loss between truth and prediction applied to each output pixel-wise. As only 200 BSDS images are annotated with ground-truth figure/ground [3], we use 150 for training and 50 for testing.

Figure 1 shows results on some examples from our 50 image test set. Compared to the previous attempt [5] to use Angular Embedding as an inference engine for figure/ground, our results are strikingly better; improvement is visually apparent on every example. Quantitative evaluation corroborates this view. Benchmarking boundary ownership prediction, our system achieves 69% accuracy compared to 58% for [5].

Though trained only on the BSDS, our system generalizes well to other datasets. It captures layering that respects scene structure and, while having only been tuned for perceptual organization, often behaves like an object detector by popping out coherent foreground regions. Please see the full paper for these additional results.

Our work demonstrates that Angular Embedding, acting on CNN predictions about pairwise pixel relationships, provides a powerful framework for segmentation and figure/ground organization. It is the first system to formulate a robust interface between these two components. Our results are a dramatic improvement over prior attempts to use spectral methods for figure/ground organization.

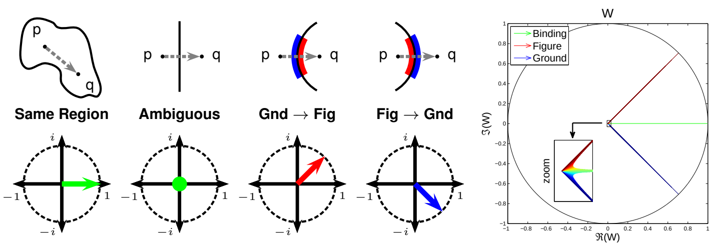

Figure 4: **Complex affinities for grouping and figure/ground.** Four basic interaction types span the space of possible pairwise pixel relationships: contiguous region, ambiguous boundary, figure → ground, and ground → figure transitions. ***Left:*** Each is captured by a single complex number, with confidence as magnitude and relative figure/ground displacement as angle from the positive real axis. ***Right:*** Combining these base cases, we express generalized affinity W as the sum of a binding force acting along the positive real axis, and figure and ground displacement forces acting at angles. W varies smoothly across its configuration space, yet exhibits distinct modes.

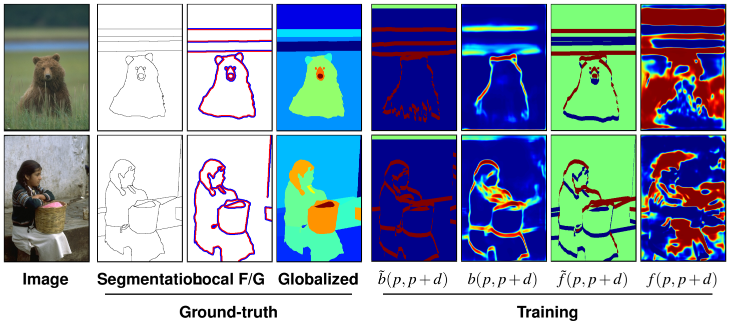

Figure 5: **Affinity learning.** ***Left:*** Given only ground-truth segmentation [7] and *local* boundary ownership [3], we infer a global ground-truth figure/ground order by running Angular Embedding with pairwise interactions defined by the local ownership. ***Right:*** Ground-truth segmentation serves to train pairwise grouping probability $b(\cdot,\cdot)$, while globalized ground-truth figure/ground trains $f(\cdot,\cdot)$. Shown are training targets $\tilde{b}$, $\tilde{f}$, and model predictions b, f, for one stencil component: the relationship between pixel p and its neighbor at relative offset $d = (-16,0)$. Ground-truth $\tilde{b}$ is binary (blue=0, red=1). $\tilde{f}$ is also binary, except pixel pairs in the same region (shown green) are ignored. As f is masked by b at test time, we require only that $f(p,q)$ be correct when $b(p,q)$ is close to 1.

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NIPS*, 2014.

[3] Charless Fowlkes, David Martin, and Jitendra Malik. Local figure/ground cues are valid for natural images. *Journal of Vision*, 2007.

[4] A. Krizhevsky, S.Ilya, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *NIPS*, 2012.

[5] Michael Maire. Simultaneous segmentation and figure/ground organization using angular embedding. *ECCV*, 2010.

[6] Michael Maire and Stella X. Yu. Progressive multigrid eigensolvers for multiscale spectral segmentation. *ICCV*, 2013.

[7] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *ICCV*, 2001.

[8] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[9] Stella X. Yu. Angular embedding: A robust quadratic criterion. *PAMI*, 2012.