

Fully connected CRF's

D.A. Forsyth

Non-local means

- Smoothing
 - Estimate the value of a pixel using pixels that are nearby
 - eg gaussian filter, etc.
 - problem: some pixels might be on the other side of an edge
- Non-local means
 - Estimate the value of a pixel using pixels that are “similar”
 - eg write pixel value v ; feature vector at pixel f ; smoothed s
 - average all pixels, weighting by similarity
 - natural questions:
 - what is f ?
 - what is w ?
 - more important:
 - how to get the sum

$$s_i = \sum_j w(\mathbf{f}_i, \mathbf{f}_j) v_j$$

Natural choices

- In

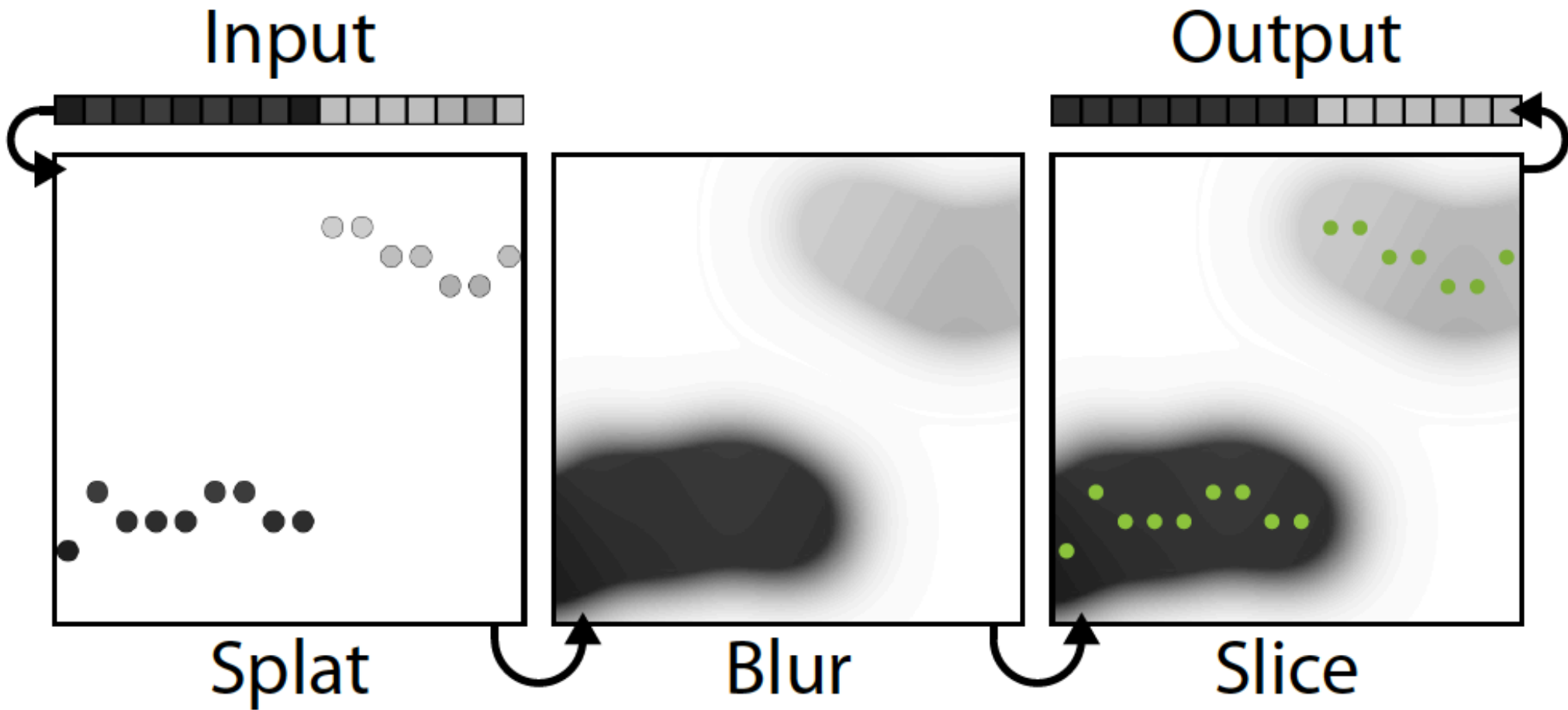
$$s_i = \sum_j w(\mathbf{f}_i, \mathbf{f}_j) v_j$$

- \mathbf{f} is
 - color, position, perhaps a texture feature
- w is

$$w(\mathbf{f}_i, \mathbf{f}_j) = \exp -\frac{1}{2} \left[(\mathbf{f}_i - \mathbf{f}_j)^T \mathcal{M} (\mathbf{f}_i - \mathbf{f}_j) \right]$$

- Notice this should simplify computing the sum
 - only “similar” pixels make reasonable contributions
 - but we must find them

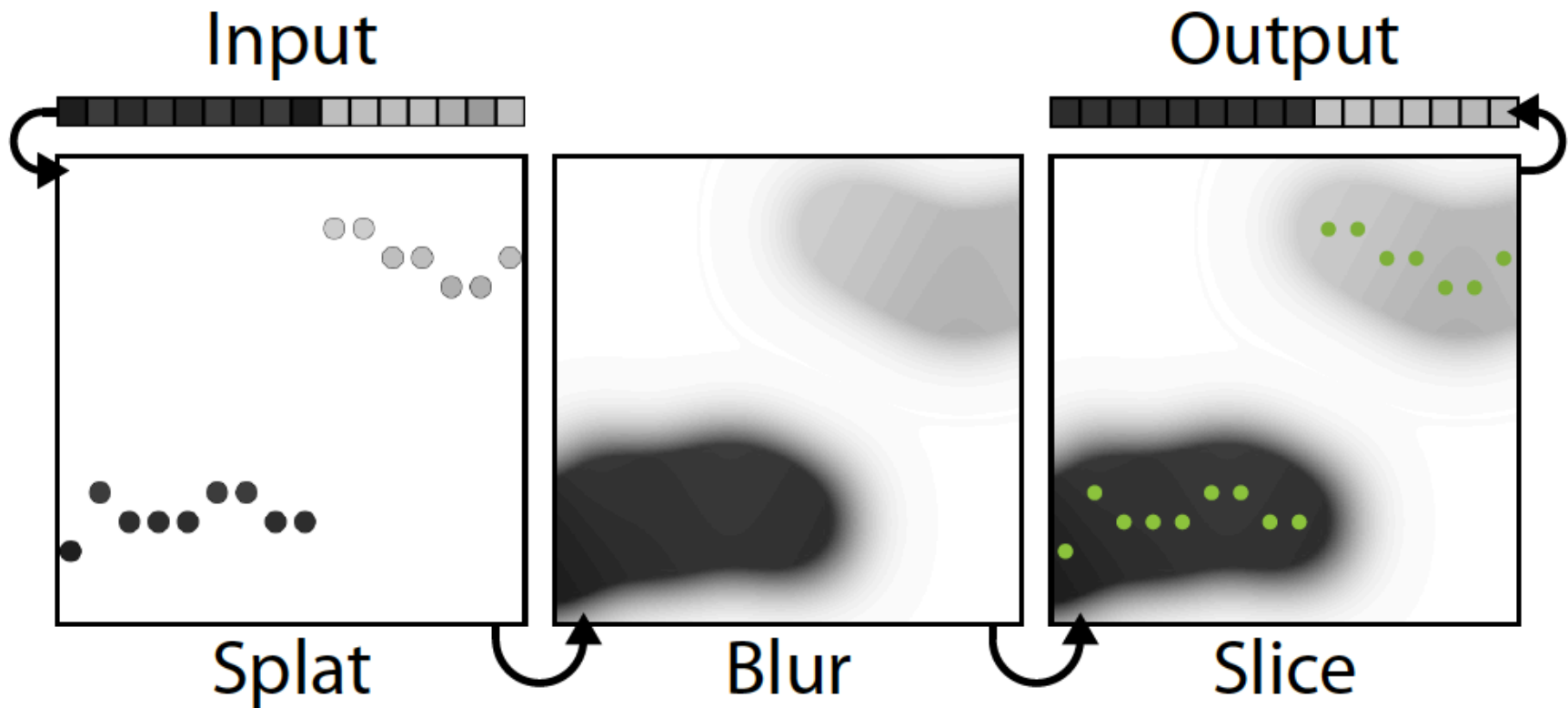
Splat, smooth, slice



from Adams, Baek, Davis

Need

- Some form of grid in high-D for f to splat onto
 - smoothing on this grid should be easy
 - it should be easy for pixels to find the closest point(s) on grid



In “high” dimension

- Permutohedral lattice

The vertices of the simplex containing any point in H_d can be computed in $O(d^2)$ time. This property will be useful for the splat and slice stages of filtering.

The nearest neighbors of a lattice point can be computed in $O(d^2)$ time. This property will be useful during the blur stage of filtering.

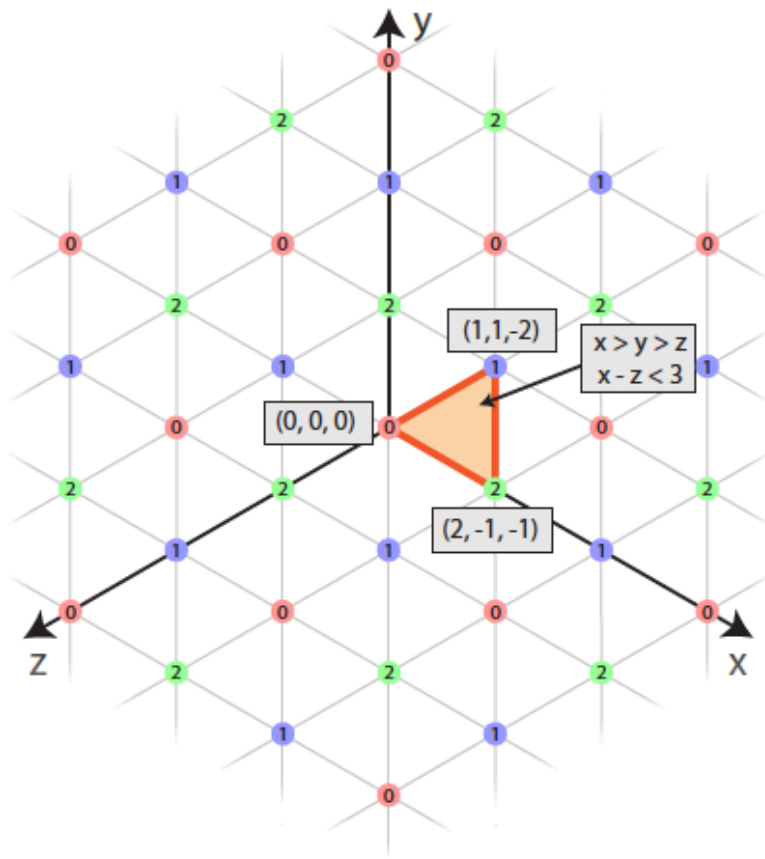


Figure 2: The d -dimensional permutohedral lattice is formed by projecting the scaled grid $(d+1)\mathbb{Z}^{d+1}$ onto the plane $\vec{x} \cdot \vec{1} = 0$. This forms the lattice $(d+1)A_d^*$, which we term the permutohedral lattice, as it describes how to tile space with permutohedra. Lattice points have integer coordinates with a consistent remainder modulo $d+1$. In the diagram above, which illustrates the case $d=2$, points are labeled and colored according to their remainder. The lattice tessellates the plane with uniform simplices, each simplex having one vertex of each remainder. The simplices are all translations and permutations of the canonical simplex (highlighted), which is defined by the inequalities $x_0 > x_1 > \dots > x_d$ and $x_0 - x_d < d+1$.

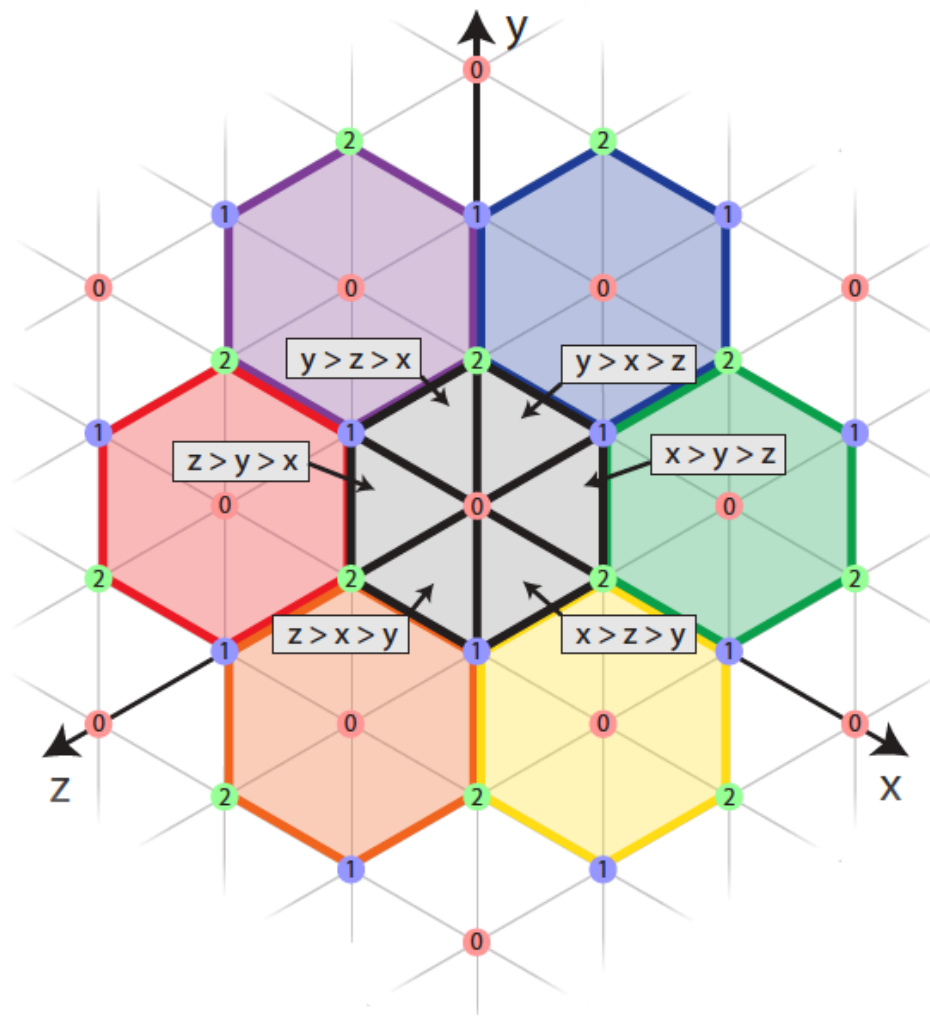


Figure 3: When using the permutohedral lattice to tessellate the subspace H_d , any point $\vec{x} \in H_d$ is enclosed by a simplex uniquely identified by the nearest remainder-0 lattice point \vec{l}_0 (the zeroes highlighted in red) and the ordering of the coordinates of $\vec{x} - \vec{l}_0$. The nearest remainder-0 lattice point can be computed with a simple rounding algorithm, and so identifying the enclosing simplex of any point and enumerating its vertices is computationally cheap ($O(d^2)$).

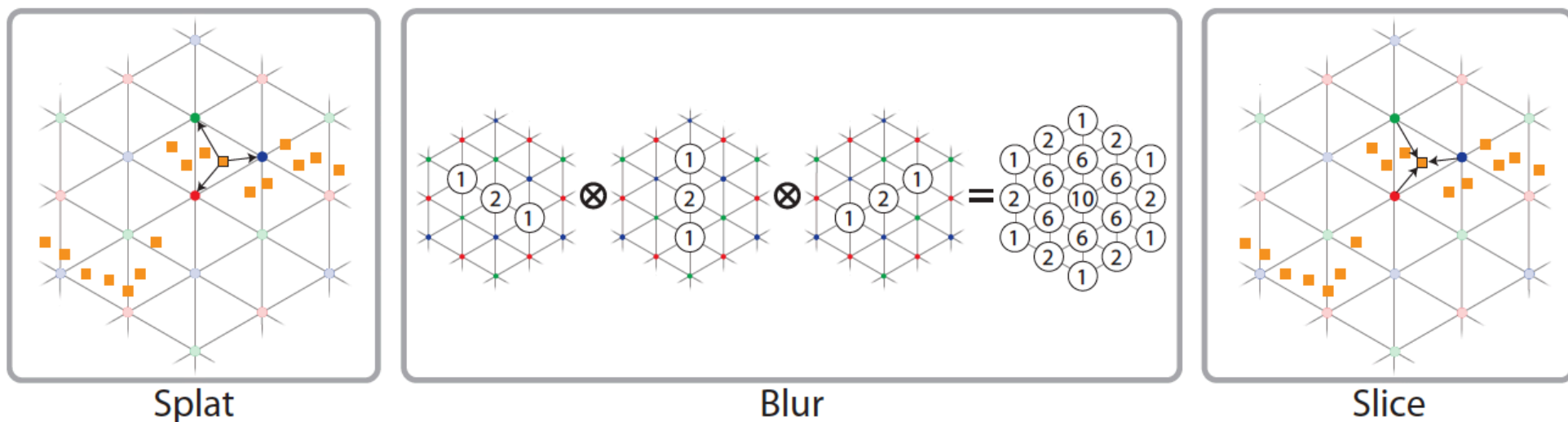


Figure 4: To perform a high-dimensional Gaussian filter using the permutohedral lattice, first the position vectors $\vec{p}_i \in \mathbb{R}^d$ are embedded in the hyperplane H_d using an orthogonal basis for H_d (not pictured). Then, each input value **splats** onto the vertices of its enclosing simplex using barycentric weights. Next, lattice points **blur** their values with nearby lattice points using a separable filter. Finally, the space is **sliced** at each input position using the same barycentric weights to interpolate output values.

Important case

- Imagine

$$w(\mathbf{f}_i, \mathbf{f}_j) = \exp -\frac{1}{2} \left[(\mathbf{f}_i - \mathbf{f}_j)^T (\mathbf{f}_i - \mathbf{f}_j) \right]$$

- then we could do each dimension separately, and multiply

Q: why not...

- do non-local-means smoothing
 - directly on semantic segmenter feature maps
- A: never seen it, don't know why
- possible A: FCCRF (next) is better?
- possible A: doesn't respect labels?

Fully connected CRF

In the fully connected pairwise CRF model, \mathcal{G} is the complete graph on \mathbf{X} and $\mathcal{C}_{\mathcal{G}}$ is the set of all unary and pairwise cliques. The corresponding Gibbs energy is

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j), \quad (1)$$

where i and j range from 1 to N . The unary potential $\psi_u(x_i)$ is computed independently for each pixel by a classifier that produces a distribution over the label assignment x_i given image features. The unary potential used in our implementation incorporates shape, texture, location, and color descriptors and is described in Section 5. Since the output of the unary classifier for each pixel is produced independently from the outputs of the classifiers for other pixels, the MAP labeling produced by the unary classifiers alone is generally noisy and inconsistent, as shown in Figure 1(b).

Why bother?

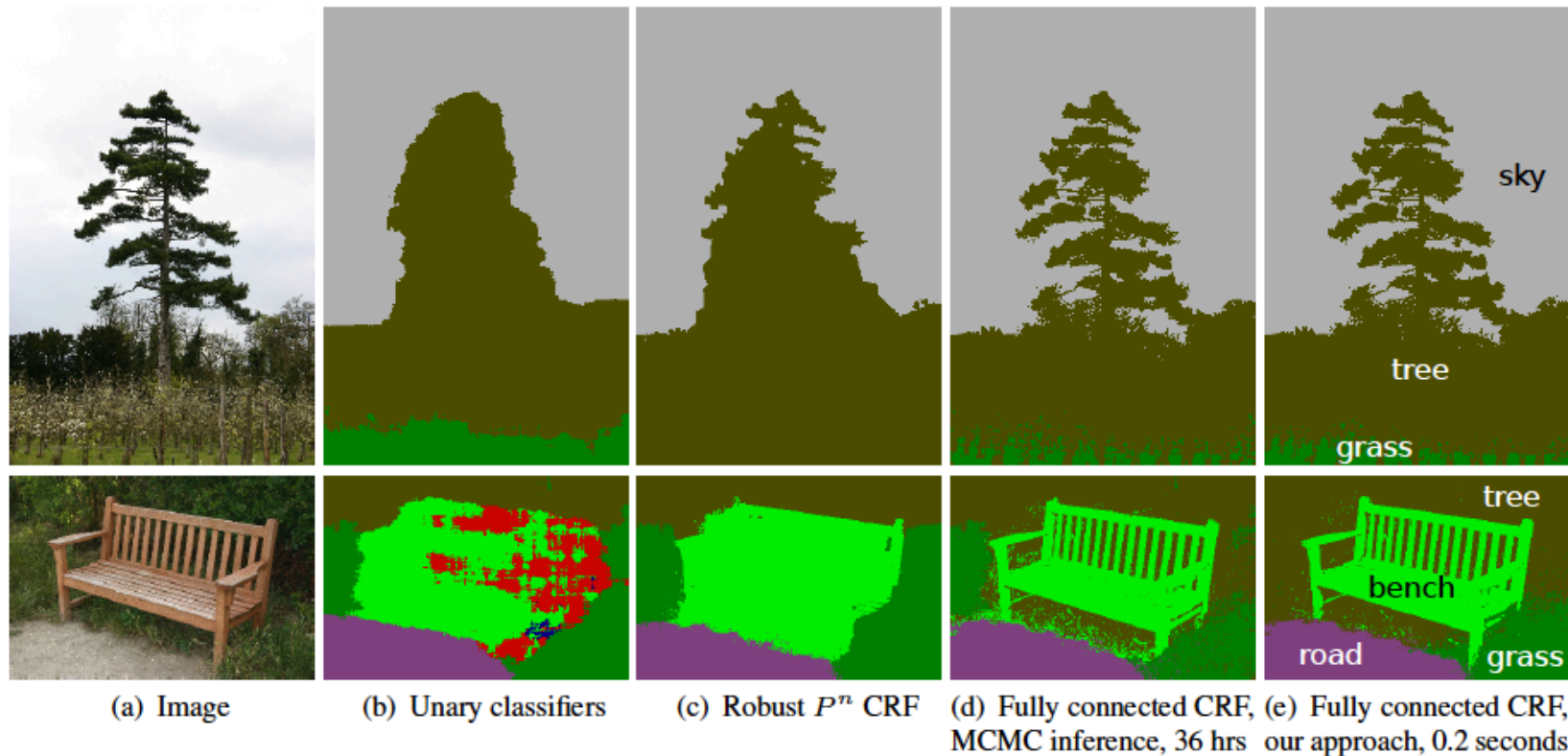


Figure 1: Pixel-level classification with a fully connected CRF. (a) Input image from the MSRC-21 dataset. (b) The response of unary classifiers used by our models. (c) Classification produced by the Robust P^n CRF [9]. (d) Classification produced by MCMC inference [17] in a fully connected pixel-level CRF model; the algorithm was run for 36 hours and only partially converged for the bottom image. (e) Classification produced by our inference algorithm in the fully connected model in 0.2 seconds.

where i and j range from 1 to N . The unary potential $\psi_u(x_i)$ is computed independently for each pixel by a classifier that produces a distribution over the label assignment x_i given image features. The unary potential used in our implementation incorporates shape, texture, location, and color descriptors and is described in Section 5. Since the output of the unary classifier for each pixel is produced independently from the outputs of the classifiers for other pixels, the MAP labeling produced by the unary classifiers alone is generally noisy and inconsistent, as shown in Figure 1(b).

The pairwise potentials in our model have the form

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \underbrace{\sum_{m=1}^K w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j)}_{k(\mathbf{f}_i, \mathbf{f}_j)}, \quad (2)$$

where each $k^{(m)}$ is a Gaussian kernel $k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) = \exp(-\frac{1}{2}(\mathbf{f}_i - \mathbf{f}_j)^T \Lambda^{(m)}(\mathbf{f}_i - \mathbf{f}_j))$, the vectors \mathbf{f}_i and \mathbf{f}_j are feature vectors for pixels i and j in an arbitrary feature space, $w^{(m)}$ are linear combination weights, and μ is a label compatibility function. Each kernel $k^{(m)}$ is characterized by a symmetric, positive-definite precision matrix $\Lambda^{(m)}$, which defines its shape.

For multi-class image segmentation and labeling we use contrast-sensitive two-kernel potentials, defined in terms of the color vectors I_i and I_j and positions p_i and p_j :

$$k(\mathbf{f}_i, \mathbf{f}_j) = \underbrace{w^{(1)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + \underbrace{w^{(2)} \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}. \quad (3)$$

The *appearance kernel* is inspired by the observation that nearby pixels with similar color are likely to be in the same class. The degrees of nearness and similarity are controlled by parameters θ_α and θ_β . The *smoothness kernel* removes small isolated regions [19]. The parameters are learned from data, as described in Section 4.

Warm-up: CRFs + variational inference

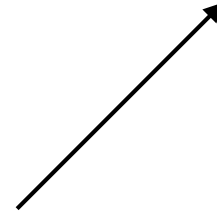
Variational Inference for FCCRFs

Minimizing the KL-divergence, while constraining $Q(\mathbf{X})$ and $Q_i(X_i)$ to be valid distributions, yields the following iterative update equation:

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^K w^{(m)} \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l') \right\}. \quad (4)$$

A detailed derivation of Equation 4 is given in the supplementary material. This update equation leads to the following inference algorithm:

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp \left\{ -\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \mu(l, l') \sum_{m=1}^K w^{(m)} \sum_{j \neq i} \underline{k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l')} \right\}$$



This should look like non-local means to you

Alg.

non-local means

Algorithm 1 Mean field in fully connected CRFs

Initialize Q

while not converged **do**

$$\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l) \text{ for all } m$$

$$\hat{Q}_i(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$$

$$Q_i(x_i) \leftarrow \exp\{-\psi_u(x_i) - \hat{Q}_i(x_i)\}$$

normalize $Q_i(x_i)$

end while

$$\triangleright Q_i(x_i) \leftarrow \frac{1}{Z_i} \exp\{-\phi_u(x_i)\}$$

\triangleright See Section 6 for convergence analysis

\triangleright **Message passing** from all X_j to all X_i

\triangleright **Compatibility transform**

\triangleright **Local update**

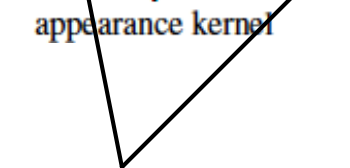
This is a vector of values, one per label l , hence the notation issue

This is a vector of values, one per label l , hence the notation issue

Improvements

- Map f to an independent basis
 - do this by
 - estimating covariance of f
 - whitening
- Now each dimension is independent, and nlm is easier

Long range connections seem to help

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + w^{(2)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}.$$


Manipulate these

Long range connections seem to help

Long-range connections. We have examined the value of long-range connections in our model by varying the spatial and color ranges θ_α and θ_β of the appearance kernel and analyzing the resulting classification accuracy. For this experiment, $w^{(1)}$ was held constant and $w^{(2)}$ was set to 0. The results are shown in Figure 6. Accuracy steadily increases as longer-range connections are added, peaking at spatial standard deviation of $\theta_\alpha = 61$ pixels and color standard deviation $\theta_\beta = 11$. At this setting, more than 50% of the pairwise potential energy in the model was assigned to edges of length 35 pixels or higher. However, long-range connections can also propagate misleading information, as shown in Figure 7.

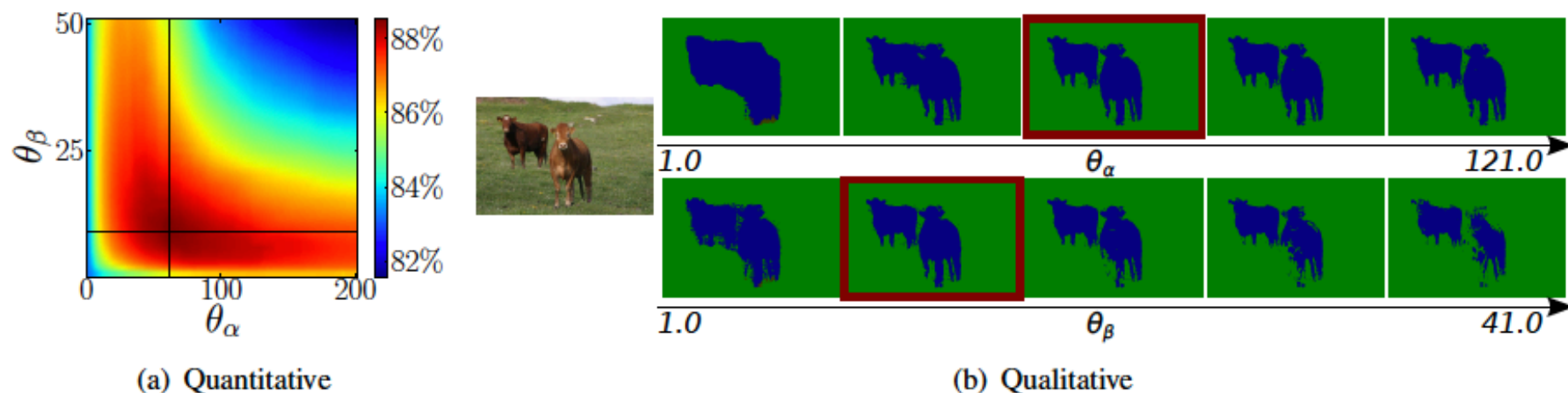
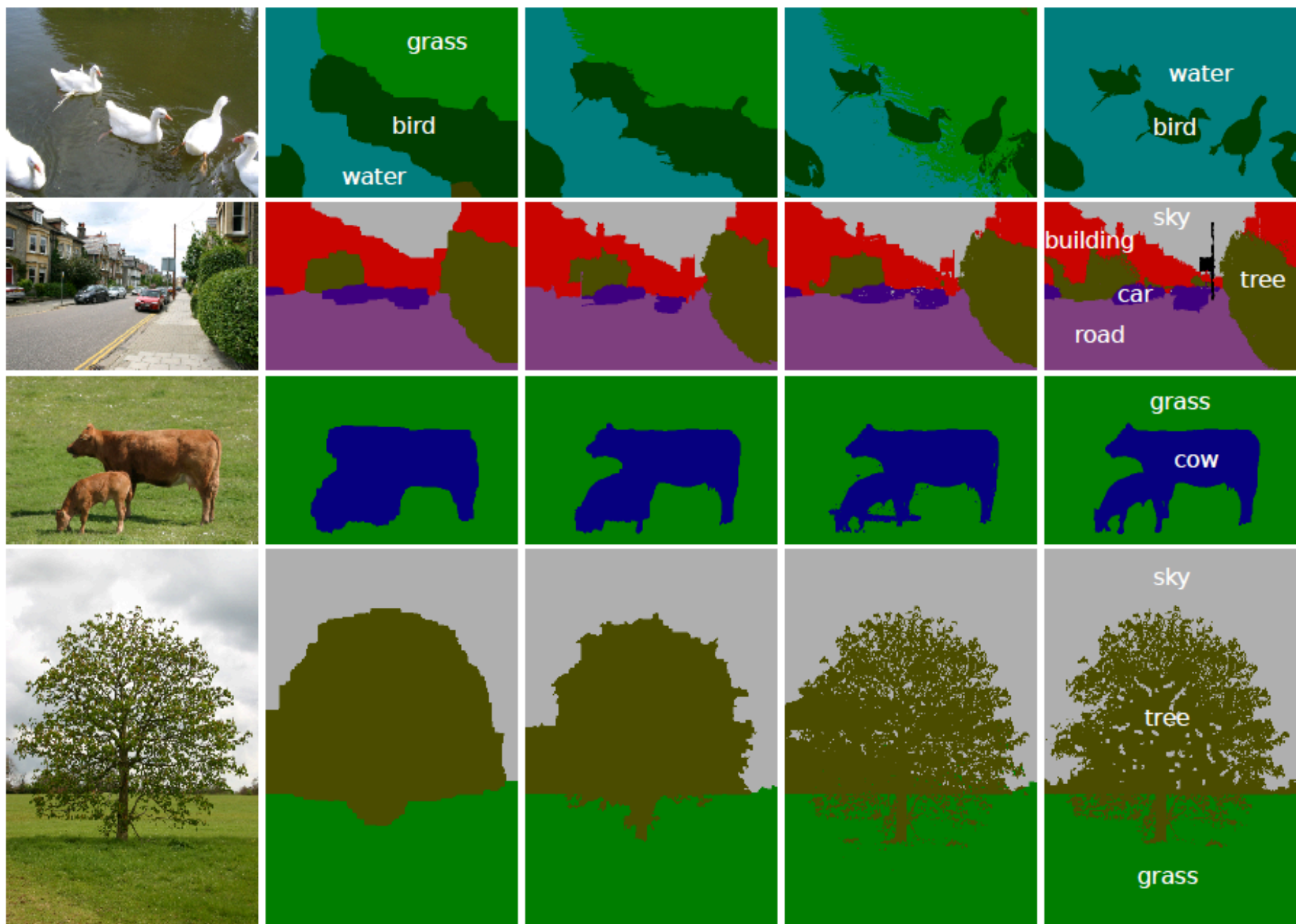


Figure 6: Influence of long-range connections on classification accuracy. (a) Global classification accuracy on the 94 MSRC images with accurate ground truth, as a function of kernel parameters θ_α and θ_β . (b) Results for one image across two slices in parameter space, shown as black lines in (a).



Image

Grid CRF

Robust P^n CRF

Our approach

Accurate ground truth

General summary

- Complicated but fast and efficient method
 - imposes spatial priors
 - results are pre deep learning
 - no end-to-end training
- For a while, widely used on semantic segmenters
 - train segmenter end-to-end
 - then bolt this on to smooth labels
 - now somewhat less common
 - why? not sure
- Weight training method exists
 - essentially, search