UFo - Coupling Uncertain Active Constellation Models with Cascaded Forest Predictors for Sematic Segmentation

Anonymous CVPR submission

Paper ID ****

Abstract

We consider the task of model-based semantic segmentation. The model is described by a constellation model of parts that are represented by active shape- and appearance models. We term this an active constellation model. As a running example we utilize a 21-part, chain-based spine model of a zebra fish observed in microscopic images. The prevailing approach to solve this task is to first generate pixel-independent features for each part, e.g. via a cascaded decision forest predictor, which are then fed into an MRF-based model-fitting objective to infer the optimal MAP solution of the constellation model. Our key contribution is to abandon this static, two-stage approach and mix feature generation and model-based inference in a new, more flexible, way. In particular we interleave the cascaded forest predictors with inference steps for the model-fitting. A key finding is that uncertain model-outputs at intermediate stages of the cascade, in the form of part-based marginals, are essential for best performance. This is because, as opposed to MAP inference, the soft marginals do not commit to a certain – potentially wrong – solution "at first sight". If unsure at first sight, soft marginals allow for "narrowing down" on the correct solution in later stages of the cascade. We validate our findings with an in-depth study of alternative inference steps, including popular geodesic smoothing as well as MAP inference. We believe that our findings are not only relevant for other types of constellation models, but, more generally, for the recent trend of combing deep learning models with physically-motivated structured models.

1. Introduction

Many tasks in computer vision have as input an image and as output a dense labeling, where each pixel is assigned one out of many pre-defined classes. An example is a sematic segmentation of a person in an image, where each pixel is assigned a label such as background, left leg, or head. So-called structured models, such as a Conditional Random Field (CRFs), are often used for semantic segmentation. Depending on the prior knowledge about the task at hand, the underlying graph may be a (super-)pixel grid, or a graphical constellation model that captures relative locations of multiple parts of an object.

Such structured models commonly capture the task of semantic segmentation via an objective function that is composed of a data term and a prior. The data term, referred to as "features", is derived from the image at hand, yielding pixel-wise distributions over class labels. The prior is enforced subsequently. Current state-of-the-art approaches typically employ pixel-wise classifiers combined with MAP inference on a (super-)pixel grid graph for semantic segmentation (e.g. [7, **?**]), or on a graphical constellation model for the localization of object parts (e.g. [8, 9, 5, 22]). A recent trend in computer vision is to learn deep, *cascaded* models for feature generation, such as CNNs [14] (see also e.g. [7]) and Auto-Context Models [23] (see also e.g. [20]). These models play the role of learning a complex non-linear mapping from images to features which are relevant for the task at hand.

This modelling framework is however very static, as it separates feature generation and inference (i.e. "'model fitting"'). It has been shown that better features can be generated by *interleaving* feature generation with MAP inference in pixel-grid structured models [19, 10, 21] or model agnostic smoothing smoothing [13].[1]

In this work we take the idea of interleaving feature generation and inference a step further: Instead of interleaving feature generation with a pixel-level structured model or model-agnostic smoothing, we interleave with a global, generative *active constellation model*. By this term we refer to a graphical constellation model, where the individual object parts are captured by *active appearance models* [4]. We suggest a cascaded pipeline, as illustrated in Figure 1. The most important aspect of this cascade is the question of what

[1] Note that this is different from the classical "'hierarchical"' approach that, purely for the sake of run-time, performs feature generation and inference multiple times on different scales (see e.g. [3, 15] [check!]).

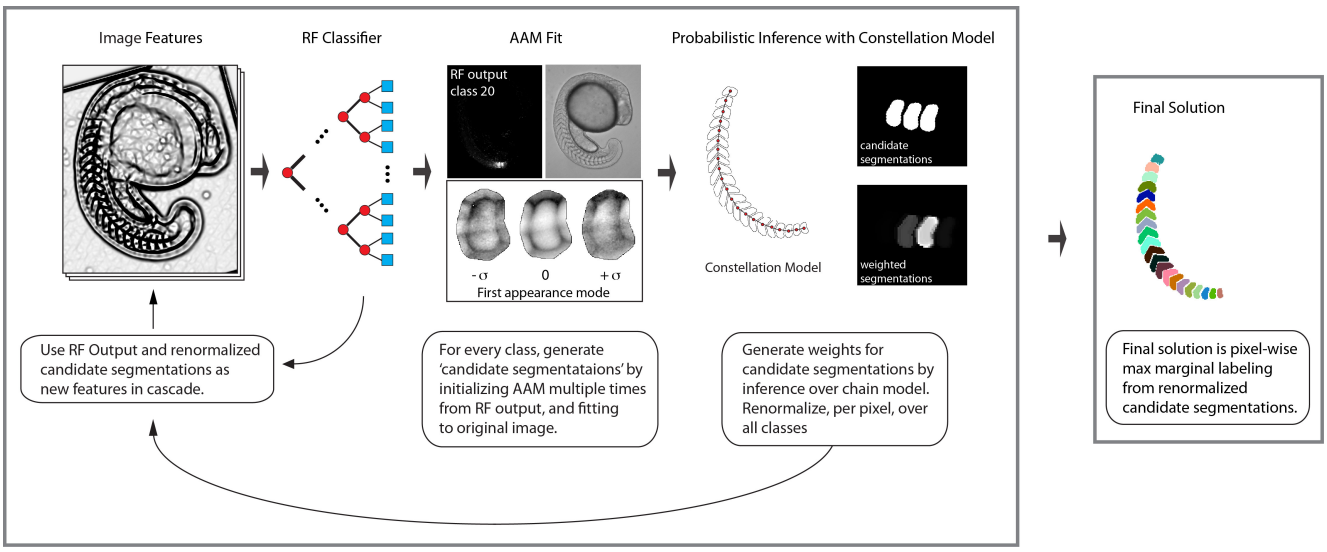

Figure 1. Pipeline.

to infer from the constellation model at intermediate stages of the cascade. Options are marginal distributions or the MAP solution. A model-agnostic, yet "image-aware" alternative to model-based inference is geodesic smoothing [13]. One of the main aspects of this work is to study the trade-offs that come with these options.

We show that marginal distributions are a clear winner for an exemplary application of semantic segmentation of many self-similar structures, namely vertebrae in spines of zebra-fish embryos. (These vertebrae are called "somites".) The reason that *uncertainy is beneficial* here is that individual somites are highly ambiguous with respect to shape and appearance, hence only the relative spatial arrangement can disambiguate this situation. Related work has tackled semantic segmentation of vertebrae (in CT scans of humans) via the classical approach of feature generation followed by MAP inference in a constellation model [9], without cascading. We show that employing marginals instead of MAP in a cascaded feature generation pipeline help to avoid committing to a wrong solution in the early stages of a cascade, and lead to a major increase in resulting segmentation accuracy.

Closely related to the presented work are (1) Auto Context [23], but they do not perform any smoothing in between levels of the cascade. (2) Geodesic Forests [13], but they do not use a generative structured model for smoothing. (3) Cascaded classifiers interleaved with MAP inference [], but they do not use a global generative model and do not explore marginals for inference. (4) Constellation models for vertebrae and other self-similar object segmentation [8, 9, 12, 5, 11] but they do not run a cascade, and do not exploit marginal distributions.

To summarize, we claim the following three **contributions**:

- In the field of semantic segmentation with constellation models, we are the first to interleave feature generation and model-based inference. We show that this boosts performance considerably, compared to not cascading.
- We show, for the first time, that probabilistic inference gives a major (6%) boost in performance in cascaded MRF-Forest-based models. This is compared to standard MAP inference (as e.g. in [9, 22, 5]) and model-agnostic geodesic smoothing [13, 18].
- We are the first to tackle spine detection in zebra fish, where we achieve an overall average Dice score of 82%.

TODO: Additional Literature: Deformable Templates Guided Discriminative Models for Robust 3D Brain MRI Segmentation. [16] Liu et al. 2013: Uses generative model to "refine" features in a cascaded discriminative classifier. Uwe's hint [6]: As Time Goes by – Anytime Semantic Segmentation with Iterative Context Forests

2. Background

Random Forests and Cascading. We assume that the reader is familiar with the general concept of Random Forests (RF) [1]. In the Auto Context approach [23], the probability maps yielded by a random forest are fed as features into subsequent random forests, yielding a cascade of forests. Interleaved inference/smoothing operates on these probability maps, and feeds "smoothed" versions of them into the next forest [19, 10, 21, 13].

AAM Active appearance models (AAMs) [4] are linear, generative, parametric models of shape and appearance that are learned from training data, and are widely used e.g., for face modelling (XX need ref). Fitting an AAM is a non-linear optimisation problem that consists of finding the model instance that minimizes the error to the input image. Optimization is commonly done either by learning a linear mapping from the error image to parameter updates [4], or by iteratively computing incremental gradient descent updates to model parameters [17].

The Shape Model is defined as follows:

$$s = s_0 + \sum_{i=1}^{n} p_i s_i$$

where s is a vector of x,y coordinates of the landmarks that define the shape. From PCA, s_0 is the mean shape, and s_i are n eigenvectors corresponding to the n largest eigenvalues. Not shown here is that the training data is first noramalised using a Procrustes analysis with a global shape normalising transformation (in our case, a similarity transform) to avoid modeling this variation in the shape model.

The Appearance Model is defined on the base-mesh, as follows:

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$$

Importantly, A_0 and A_i are computed from PCA on a set of *shape normalised* training images, which have been warped onto the base-mesh x, which is defined by the mean shape s_0. Shape normalization is a key benefit of AAMs (compared to e.g. Eigen-faces), and leads to more compressed PCA representation. XX Additionally, an even more compact representation can be realized by a subsequent step of PCA on the combined shape and appearance parameters, leading to a Combined AAM; however, this limits the choice of efficient solvers, such as the Inverse Compositional Algorithm (XX cite earlier Baker paper). For the rest of this paper, we will restrict ourselves to discussing independent shape and appearance models.

Since AAMs are generative models, they can be used to create model instances which can be directly compared to the input image. Thus, fitting an AAM consists of finding the model parameters that minimize the sum-of-squared-distances between the image and the corresponding model instance, evaluated on the base mesh.

$$Cost = \sum_{x \epsilon s_0} [A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(N(W(x;p);q))]^2$$

To create a model instance, first render an image A(x) (defined by λ) on the base-mesh, and then warp it to s (defined by p). Warping can be done using e.g., a piecewise affine warping over a triangulated mesh, or a thin-plate spline, parameterized by the set of landmarks, s_0 and s. This defines the unique warp parameterised by p, called W(x;p). Finally, the model instance is transformed into the image by the global shape normalising transform N(x;q), in our case a similarity transform, parameterized by q.

A central challenge of AAMs is their sensitivity during the fitting process. A popular strategy for combatting this is to add priors on the model parameters, as follows:

$$\sum_{x \epsilon s_0} [A(x) - I(N(W(x;p);q)]^2 + \sum_{i=1}^{K} F_i^2(p,q)$$

Fortunately, AAMs can be easily extended to include priors on the model parameters (see also [17]), with minimal additional cost to the fitting algorithm, and can be interpreted in the Bayesian framework as Gaussian Regularization.

3. Method

Given an image as input, we seek a pixel-wise multi-class labelling as output, i.e. a *semantic segmentation*. We assume that a model of the spatial relation of classes can be learned, i.e. a *constellation model*. This is the case for many applications, as e.g. body part segmentation in natural [20] or medical images [22], vertebra segmentation [8, 9], etc.

We propose the following pipeline for *model-based semantic segmentation*, as layed out in Figure 1: First, we generate probability maps for each class with a random forest classifier. Second, we generate many *part proposals* for each class, with the help of Active Appearance models (cf. Sec. 3.1). Each part proposal is a binary segmentation of the respective class. It serves as a "segmentation hypothesis". Third, we perform probabilistic inference in a constellation model to weigh part proposals (cf. Sec. 3.2), and effectively "smooth" the probability maps generated by the RF classifier. Fourth, we feed the resulting "smoothed" probability maps, together with the original probability maps as well as all image features used as input to the previous RF, into a next RF classifier. To generate a resulting labeling per pixel from the last RF output in the cascade, one can take the class with maximum probability according to either the RF probability maps, or the respective "smoothed" versions.

3.1. Generating Part Proposals

Given an RF-generated probability map of a class, we First compute its centroid via the mean shift algorithm. Second, we fit an average constellation model (i.e. a static constellation of landmarks) to these centroids to yield an optimal global similarity transform w.r.t. the sum of squared landmark distances. In our application, this is sufficient to

define an approximate orientation of the part. Third, we sample a number of candidate locations around the centroids of the RF-generated probability maps to get sets of location initializations for the respective classes. Fourth, we fit a class specific active appearance model (AAM) to the image, multiple times, starting at the initial locations computed in the previous step. Each AAM fit results in a binary segmentation, together with a cost for the fit (cf. Eq. (**??**)). These binary segmentations serve as part proposals, i.e. segmentation hypotheses for their respective classes.

3.2. Weighting and Fusing Part Proposals

The above method generates a number of part proposals, i.e. binary segmentation hypotheses, per class. We assign weights to these proposals by means of a constellation model in the form of a second order CRF. The nodes of the CRF correspond to the classes, $c \in \{1..n_C\} =: C$, and the labels of each node correspond to the respective part proposals, $l \in \{1..n_L\} =: L$. Note that here, for the sake of notation simplicity, we assume we have the same number of proposals for each class.

The unary factors, $\phi_c(l)$, reflect the cost of the respective AAM fit, $A_c(l)$, together with the RF probability map $P_c : \Omega \rightarrow [0, 1]$, accumulated over the foreground of the respective binary segmentation, $H_{c,l} : \Omega \rightarrow \{0, 1\}$:

$$\phi_c(l) = \exp\left(-\lambda \cdot A_c(l)\right) \cdot \frac{\sum_{x,y} H_{c,l}(x, y) \cdot P_c(x, y)}{\sum_{x,y} H_{c,l}(x, y)} \quad (1)$$

A parameter λ weights the relative influence of the two terms.

The pairwise factors, $\psi_{c,b}(l, k)$, reflect the probability of relative locations of neighboring proposals. To this end, we learn the average offset between part centroids, as well as respective covariances, and assume an acccording gaussian distribution.

We compute weights for each proposal and each class by means of probabilistic inference in this CRF. In our application, the respective graphical model is a chain, and hence probabilistic inference can be performed optimally and efficiently by means of dynamic programming. Given the resulting marginals $p_c(l)$, we compute a weighted average of part proposals:

$$S_c(x, y) = \frac{1}{Z(x, y)} \cdot \sum_{l \in L} p_c(l) \cdot H_{c,l}(x, y)$$

Here, $Z(x, y)$ serves for pixel-wise re-normalization; I.e., $Z(x, y) = \sum_{c \in C} \sum_{l \in L} p_c(l) \cdot H_{c,l}(x, y)$. We call S_c a *smoothed probability map* for class c.

3.3. Parameters

... to be put into respective equations spread everywhere...

- n: # of shape modes
- m: # of appearance modes
- λ: trust in image cost vs. discr output
- n_L: # of model instances that are initialized for fitting
- g: # of grad descent steps

4. Results

We evaluate our semantic segmentation approach on a data set of 32 images of developing zebrafish. We manually segment these images into 22 classes, corresponding to 21 segments of the developing spine and background. This data set poses multiple challenges, due to the repetitive nature of the structures and the small amount of training data. We provide comparisons with the state-of-the-art approach for spine detection, and a range of other cascaded random forest classifiers, including Auto-context (XX cite), GeoF (XX cite), and MAP cascade (XX cite). We trained all algorithms with the same input features and the same forest parameters, 3 levels of 16 trees, each with depth 12. Features stem from a standard filter bank [Weka], together with local contextual features, consisting of offset features values, and their differences to the local feature value. Finally, quantitative evaluation is given in terms of the Dice score averaged over all classes of the 32 images, using 2-fold cross-validation.

Approach: Cascade interleaved with "'smoothing"'. point to figure below

Dataset: 32 light microscopic images of developing zebra-fish. Task: Semantic segmentation of 21 somites (i.e. developing vertebrae). We perform a two-fold cross validation. We explore four different approaches.

note to dave - Smoothing Figure

Figure 2 show the different types for inference/smoothing we evaluate.

Ours. AAM with probabilistic inference

Auto Context.

GeoF. The simplest way to generate a smoothed RF Output is to re-use the Geodesic smoothing idea from Criminisi. Let:

$$Q(x; M, \nabla I) = \min_{x'}(\delta(x, x') + \nu M)$$

and ν is some free parameter. Note, x and x' are two points in the image. $M(x') = 1 - p(c|v(x'))$, where v(x') is the feature vector at pixel x'. Then the smoothed RF output is calculated as:

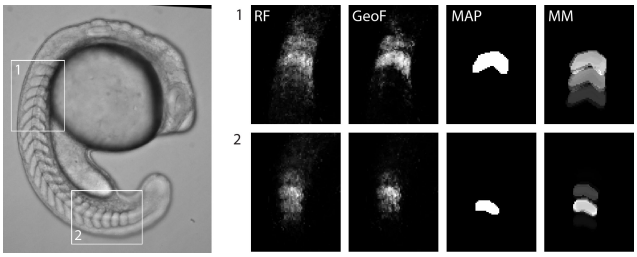

Figure 2. In our evaluation we interleave different types of inference/smoothing into our cascaded random forest pipeline for semantic segmentation of somites (i.e. vertebrae) in zebra-fish embryos. The Figure shows examples of the types of inference/smoothing we evaluate. Left: Exemplary zebra-fish embryo. Boxes (1,2): Exemplary somites. Right: Close-ups on probability maps of the respective class labels. Note that close-ups on (2) are rotated by 90 degrees. Four versions of smoothing/inference: (RF) random forest probability map; (GeoF) smoothed by geodesic smoothing[**?**]; (MAP) "'smoothed"' by MAP inference in our constellation model, yielding a binary "'probability map"'; (MM) "'smoothed"' by our proposed approach, i.e. probabilistic inference in our constellation model.

	RF Output	Smoothed RF Output
Plain Cascading (Auto-context)	0.60 (0.20)	-
Geodesic	0.63 (0.21)	0.66 (0.22)
MAP	0.71 (0.27)	0.76 (0.27)
Model Marginals	**0.82** (**0.16**)	**0.82** (0.18)

Figure 3. Evaluation on 32 datasets. Dice Scores on all 21 Somites: Mean and standard deviation (in brackets).

$$g(c|v(x)) = \frac{1}{Z} p(c|v(x)) e^{\frac{-Q(x;p(c|v(\Omega)),\nabla J)^2}{\sigma^2}}$$

This accomplishes smoothing in quite an indirect way, as a competition between different possible class labels for a given pixel, mediated by the normalization Z.

MAP Inference. talk about cascade boxplots and table (below)

5. Discussion

Ours is the best :)

Cascading Helps! ... performance goes up over levels. Boxplots!

... "'traditional"' MAP after one level sucks.

Smoothing Helps! Model-based Smoothing Helps Best! ... Auto Context $<$ GeoF $<$ Model Based. Reason: More Specific Prior knowledge...

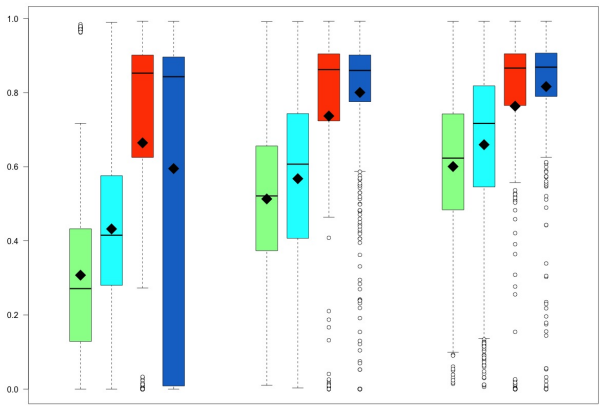

Figure 4. 3 level cascade. Segmentation accuracy of 4 methods after each level: RF output (green), GeoF (cyan), MAP (red), Ours (blue). For every method at every level, Dice scores of 21 somites in 32 images, i.e. 672 scores, are visualized as a *box plot* [2]. A colored box spans from lower to upper quartile, i.e. the inter-quartile range. I.e. 50% of the data points lie within the box. The horizontal bar within the box depicts the median. The black diamond depicts the mean. Whiskers depict the outlier-free data range. Circles depict outliers. Outliers are defined as data points beyond median $\pm$ 2 inter-quartile ranges.

Uncertainty Helps! ... Probabilistic inference considerably outperforms MAP inference (6% better). Reason: With Prob. inference, cases can be rescued if not caught after first level (i.e. "'at first site"'). Show some rescue cases.

We also tried MAP after third level. Once all cases are rescued, MAP is fine (performs equally – state numbers). BUT NOT EARLIER!

Dave, in the rescue figure, please call the bottom row "'Ours"', and maybe re-arrange rows as columns so that the figure fits into one column without being too small.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 2

[2] J. Chambers. *Graphical Methods for Data Analysis*. Chapman & Hall statistics series. Wadsworth International Group, 1983. 5

[3] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7578 of *Lecture Notes in Computer Science*, pages 278–291. Springer Berlin Heidelberg, 2012. 1

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 1, 3

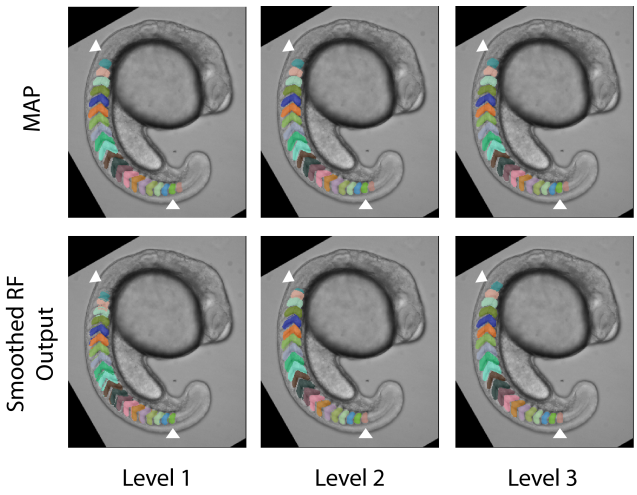

Figure 5. Rescue Case. Arrows point to ground truth start end end of spine.

[5] N. Duy, H. Lamecker, D. Kainmueller, and S. Zachow. Automatic detection and classification of teeth in ct data. In N. Ayache, H. Delingette, P. Golland, and K. Mori, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 609–616. Springer Berlin Heidelberg, 2012. 1, 2

[6] B. Fröhlich, E. Rodner, and J. Denzler. As time goes by – anytime semantic segmentation with iterative context forests. In A. Pinz, T. Pock, H. Bischof, and F. Leberl, editors, *Pattern Recognition*, volume 7476 of *Lecture Notes in Computer Science*, pages 1–10. Springer Berlin Heidelberg, 2012. 2

[7] J. Funke, J. N. Martel, S. Gerhard, B. Andres, D. C. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, H. Pfister, A. Cardona, et al. Candidate sampling for neuron reconstruction from anisotropic electron microscopy volumes. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pages 17–24. Springer, 2014. 1

[8] B. Glocker, J. Feulner, A. Criminisi, D. Haynor, and E. Konukoglu. Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans. In N. Ayache, H. Delingette, P. Golland, and K. Mori, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, volume 7512 of *Lecture Notes in Computer Science*, pages 590–598. Springer Berlin Heidelberg, 2012. 1, 2, 3

[9] B. Glocker, D. Zikic, E. Konukoglu, D. Haynor, and A. Criminisi. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, volume 8150 of *Lecture Notes in Computer Science*, pages 262–270. Springer Berlin Heidelberg, 2013. 1, 2, 3

[10] J. Jancsary, S. Nowozin, T. Sharp, and C. Rother. Regression tree fields – an efficient, non-parametric approach to image labeling problems. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2376–2383, June 2012. 1, 2

[11] D. Kainmueller, F. Jug, C. Rother, and G. Myers. Active graph matching for automatic joint segmentation and annotation of c. elegans. In P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2014*, volume 8673 of *Lecture Notes in Computer Science*, pages 81–88. Springer International Publishing, 2014. 2

[12] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. Automated model-based vertebra detection, identification, and segmentation in {CT} images. *Medical Image Analysis*, 13(3):471 – 482, 2009. 2

[13] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. Geof: Geodesic forests for learning coupled predictors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 65–72, June 2013. 1, 2

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[15] C. Lindner, S. Thiagarajah, J. Wilkinson, T. Consortium, G. Wallis, and T. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *Medical Imaging, IEEE Transactions on*, 32(8):1462–1472, Aug 2013. 1

[16] C.-Y. Liu, J. Iglesias, and Z. Tu. Deformable templates guided discriminative models for robust 3d brain mri segmentation. *Neuroinformatics*, 11(4):447–468, 2013. 2

[17] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 3

[18] A. Montillo, J. Shotton, J. Winn, J. Iglesias, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In G. Szkely and H. Hahn, editors, *Information Processing in Medical Imaging*, volume 6801 of *Lecture Notes in Computer Science*, pages 184–196. Springer Berlin Heidelberg, 2011. 2

[19] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kohli. Decision tree fields. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1668–1675, Nov 2011. 1, 2

[20] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision ECCV 2014*, volume 8690 of *Lecture Notes in Computer Science*, pages 33–47. Springer International Publishing, 2014. 1, 3

[21] U. Schmidt, C. Rother, S. Nowozin, J. Jancsary, and S. Roth. Discriminative non-blind deblurring. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 604–611, June 2013. 1, 2

[22] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu. Hierarchical parsing and semantic navigation of full body ct data. volume 7259, pages 725902–725902–8, 2009. 1, 2, 3

[23] Z. Tu. Auto-context and its application to high-level vision tasks. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. 1, 2