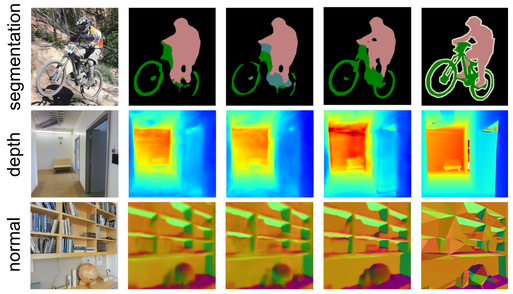
Adversarial Structure Matching for Structured Prediction Tasks

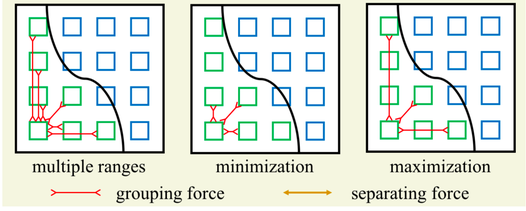
Jyh-Jing Hwang*, Tsung-Wei Ke*, Jianbo Shi, Stella X. Yu

Berkeley UNIVERSITY OF CALIFORNIA · ICSI · Penn

Overview

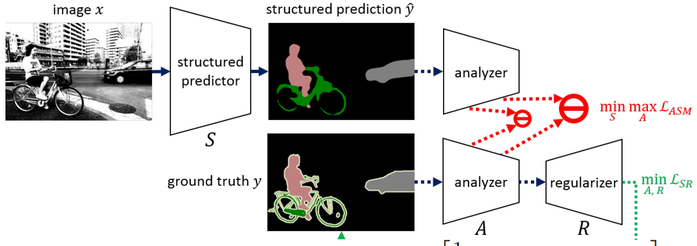


- Structured predictions output correlated 2D masks, including semantic segmentation, depth estimation, surface normal prediction, etc..
- IID (softmax / L2) is the most common approach, which ignores label correlations among pixels.
- GAN is a prior based structural model, which encodes relationships in a one-to-set mapping.
- Our ASM adversarially matches *multiscale structures* in the label space, featuring:
  1. Adaptive structure prior
  2. Instance specificity
  3. Generalizability.
- Adaptive Affinity Fields (2^{nd} order regularization): AAF only selects *pixel relationships* adaptively.



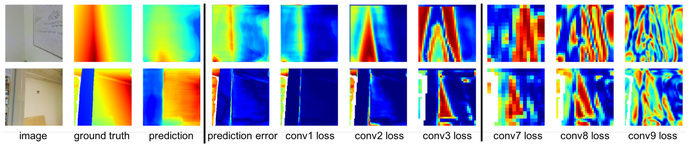
Adaptive Affinity Fields for Semantic Segmentation, ECCV 2018.

Adversarial Structure Matching

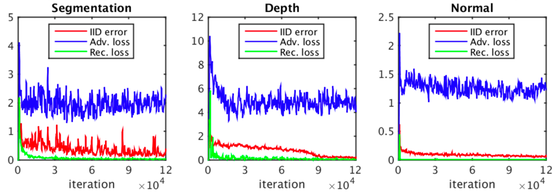


- Objective: $S^* = \underbrace{\operatorname{argmin}_S \max_A \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\frac{1}{2}\|A(S(\mathbf{x})) - A(\mathbf{y})\|_2^2\right]}_{\text{adversarial structure matching loss}} + \underbrace{\min_{A,R} \lambda \mathbb{E}_{\mathbf{y}} [IID(\mathbf{y}, R(A_t(\mathbf{y})))]}_{\text{structure regularization loss}},$

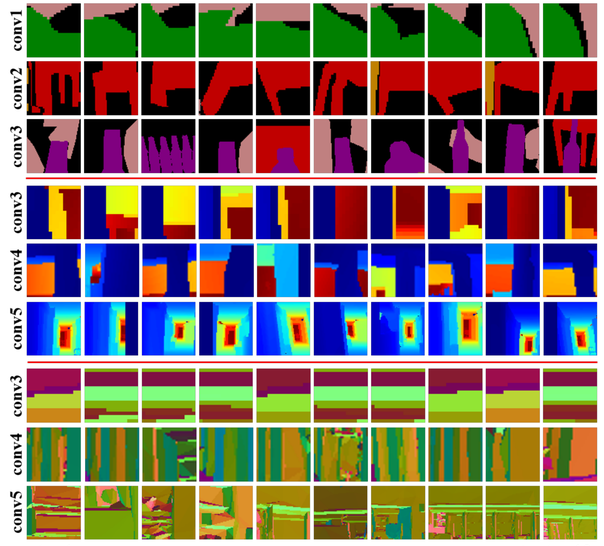
- Sampled loss patterns in different layers:



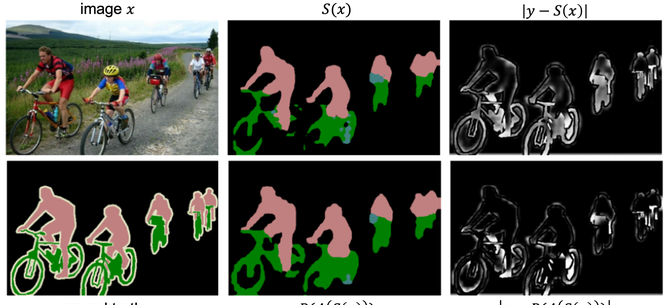
- *Analyzer* $\boldsymbol{A}$ is trained to maximize the multiscale structural mistakes of $\boldsymbol{S}$. → Hard negative mining.
- *Predictor* $\boldsymbol{S}$ is trained to minimize the same error.
- *Regularizer* $\boldsymbol{R}$ ensures that $\boldsymbol{A}$ also forms a good basis for reconstructing the ground truth.
- Assuming infinite capacity for $\boldsymbol{S}$ and $\boldsymbol{A}$, we proved $S^*(x) = y$ and $V(S^*, A^*) = 0$. (Nash equilibrium)
- ASM retains critical assessments thru training:



What's Learned in *Analyzer A*?



- $\boldsymbol{A}$ encodes multiscale pixel relationships, *e.g.*, person riding bike, hand picking up bottle, *etc.*.
- $\boldsymbol{A}$ and $\boldsymbol{R}$ learn to complete shapes while $\boldsymbol{A}$ adaptively refines the focus of supervision.



Experimental Results