Data-driven Adaptive Regularization for Imaging Problems

Byung-Woo Hong

Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea

ELTE, INFORMATIKAI KAR
Budapest, Hungary, 2023.07.17

Introduction

Composite Energy

$$\min_u \lambda \mathcal{D}(u) + (1 - \lambda) \mathcal{R}(u) \tag{1}$$

- $u : \Omega \to \mathbb{R}^N$ is an unknown function
- $\mathcal{D} : u \to \mathbb{R}$ is a data fidelity function
- $\mathcal{R} : u \to \mathbb{R}$ is a regularization function
- $\lambda \in (0, 1) \subset \mathbb{R}$ is a balancing parameter between $\mathcal{D}$ and $\mathcal{R}$

Conventional Variational Problem - Static Balancing

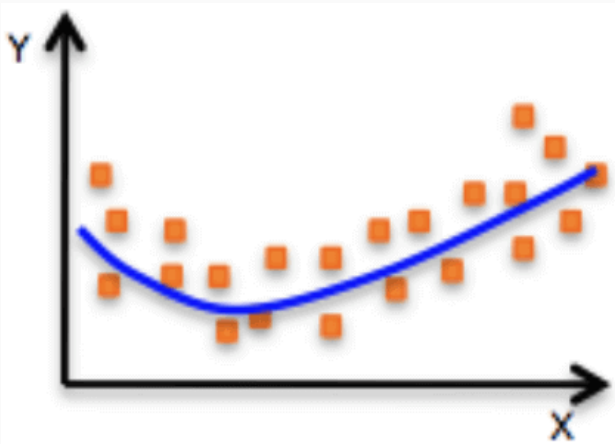
$$\min_u \lambda \mathcal{D}(u) + (1 - \lambda) \mathcal{R}(u) \tag{2}$$

$$\mathcal{D}(u) = \sum_{x \in \Omega} \rho(u(x)) \tag{3}$$

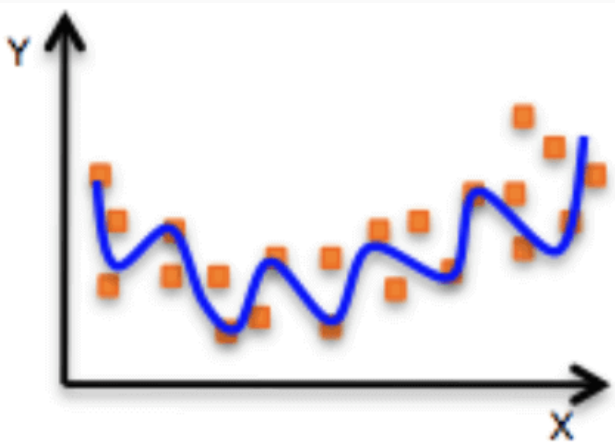
$$\mathcal{R}(u) = \sum_{x \in \Omega} \gamma(u(x)) \tag{4}$$

- $\rho : \Omega \to \mathbb{R}$ measures data fidelity
- $\gamma : \Omega \to \mathbb{R}$ measures regularity
- λ is constant in both space (domain) and time (optimization)

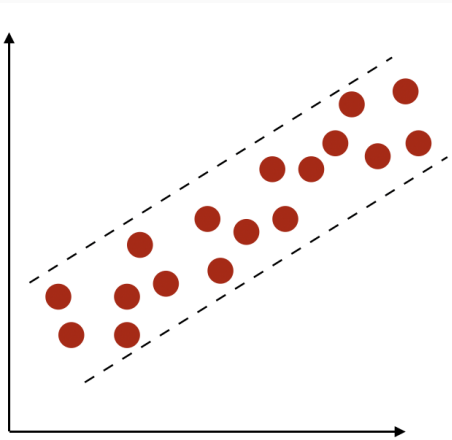
Illustrative Motivation in Regression Problem



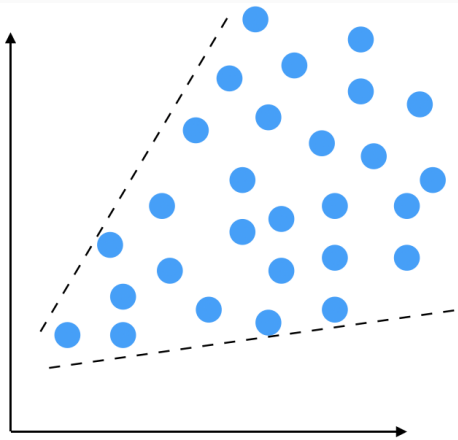
(a) robust



(b) overfitting

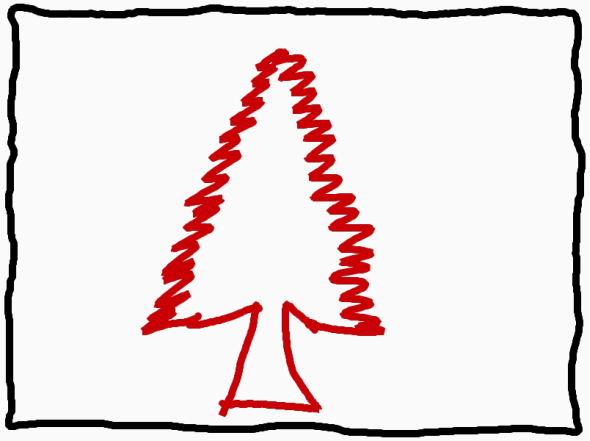


(c) homoscedasticity

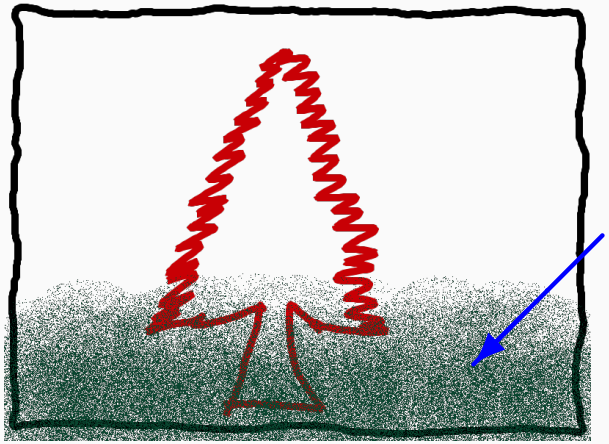


(d) heteroscedasticity

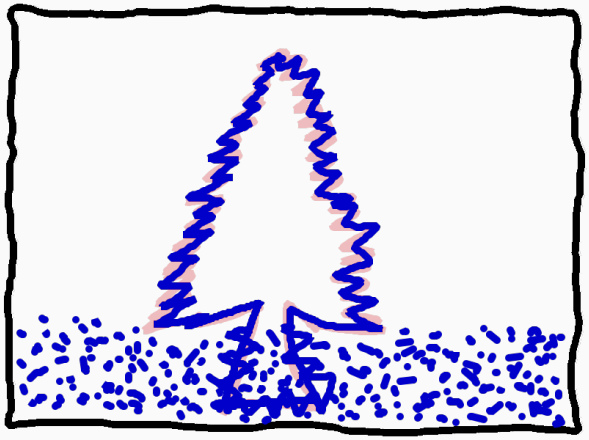
Illustrative Motivation in Image Segmentation



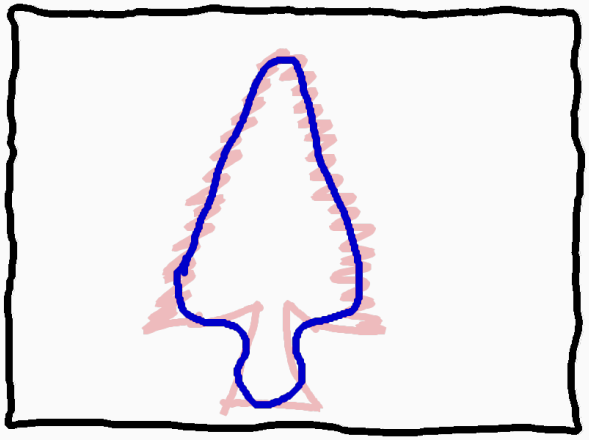
(a) image



(b) noisy image

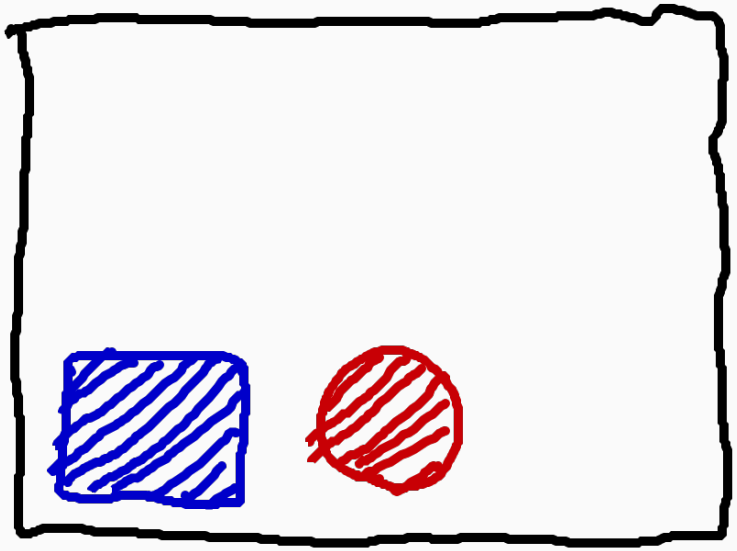


(c) small regularity

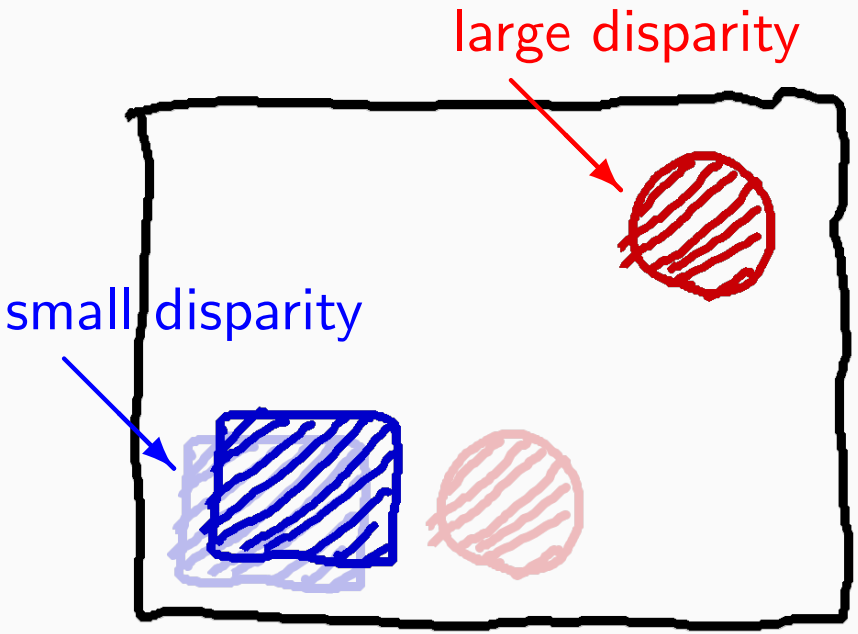


(d) large regularity

Illustrative Motivation in Motion Estimation

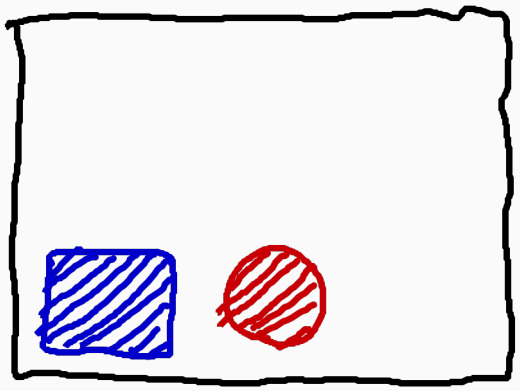


(a) first image

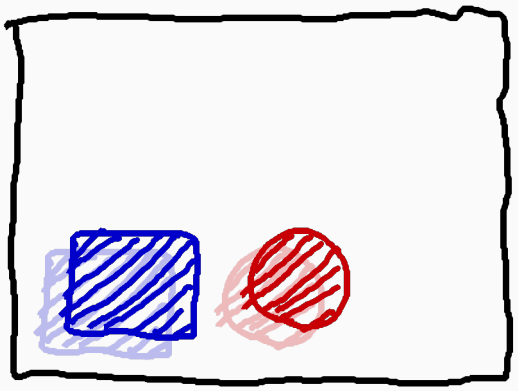


(b) second image

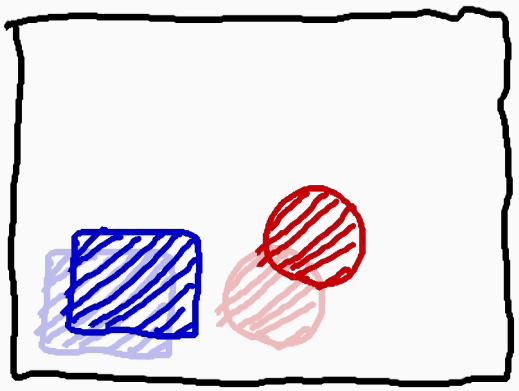
Illustrative Motivation in Motion Estimation



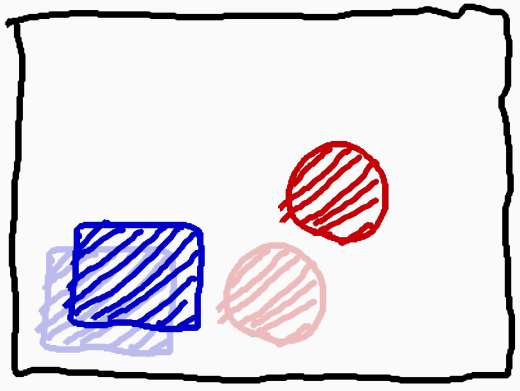
$t = 0$



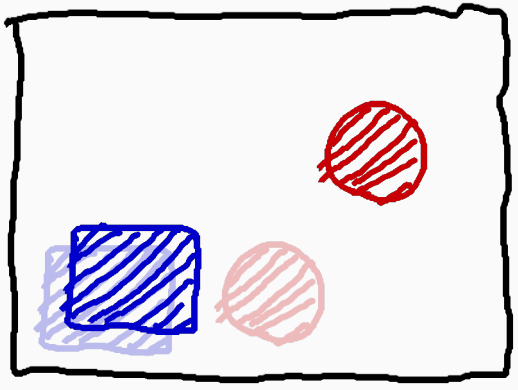
$t = 1$



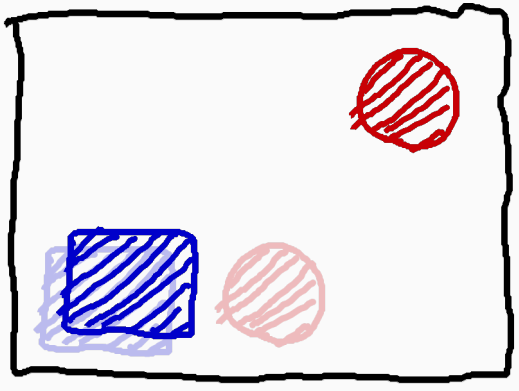
$t = 2$



$t = 3$



$t = 4$



$t = 5$

Derivation of Weight for Regularization in Denoising Problem

Additive Gaussian Noise Model

$$f = u + \eta, \quad \eta(x) \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

Bayesian Formulation

$$p(u|f) \propto p(f|u)\, p(u) \tag{6}$$

Maximum A Posteriori estimation

$$u^* = \arg\max_u p(u|f) = \arg\max_u p(f|u)\, p(u) \tag{7}$$

$$= \arg\max_u \log p(f|u) + \log p(u) \tag{8}$$

$$= \arg\min_u -\log p(f|u) - \log p(u) \tag{9}$$

Derivation of Weight for Regularization in Denoising Problem

Likelihood and Prior probabilities

$$p(f(x)|u(x)) \propto \exp\left(-\frac{1}{2\sigma^2}\int_\Omega (f(x)-u(x))^2\,\mathrm{d}x\right) \tag{10}$$

$$p(u(x)) \propto \exp\left(-\frac{1}{2\mu^2}\int_\Omega |\nabla u(x)|^2\,\mathrm{d}x\right) \tag{11}$$

Energy Functional

$$E(u) = \frac{1}{2\sigma^2}\int_\Omega (f(x)-u(x))^2\,\mathrm{d}x + \frac{1}{2\mu^2}\int_\Omega |\nabla u(x)|^2\,\mathrm{d}x \tag{12}$$

$$= \int_\Omega (f(x)-u(x))^2\,\mathrm{d}x + \boxed{\frac{\sigma^2}{\mu^2}}\int_\Omega |\nabla u(x)|^2\,\mathrm{d}x \tag{13}$$

Illustrative Motivation in Image Segmentation

(a) input image

(b) segmentation

(c) residual

(d) residual variance

Regularization via Interaction between Model and Data

- Prior works generally consider static image features such as edges only from the data in determining regularization
- It is desired to consider both the current state of the observation and the underlying model for the regularization scheme
- We propose an adaptive regularization scheme that considers both spatially and temporarily varying statistics

Adaptive Regularization

Adaptive Regularization

$$\min_u \mathcal{D}_\lambda(u) + \mathcal{R}_\lambda(u) \tag{14}$$

$$\mathcal{D}_\lambda(u) = \sum_{x \in \Omega} \lambda(u(x))\, \rho(u(x)) \tag{15}$$

$$\mathcal{R}_\lambda(u) = \sum_{x \in \Omega} (1 - \lambda(u(x)))\, \gamma(u(x)) \tag{16}$$

- $\rho : \Omega \to \mathbb{R}$ measures data fidelity
- $\gamma : \Omega \to \mathbb{R}$ measures regularity
- $\lambda : \Omega \to \mathbb{R}$ modulates ρ and γ

Bayesian criterion - Maximum A Posteriori

$$u^* = \arg\max_u p(u|f) \propto \ell(u)\, q(u) \tag{17}$$

$$\ell(u) = p(f|u) \tag{18}$$

$$q(u) = p(u) \tag{19}$$

- u is a solution for the object of interest
- f is the observation
- ℓ is a likelihood function
- q is a prior function

Bayesian criterion - Adaptive Regularization

$$u^* = \arg\max_u \ell(u)^{\lambda(u)}\, q^{(1-\lambda(u))}(u) \tag{20}$$

$$\lambda(u(x)) \propto \ell(u(x)) \tag{21}$$

- λ is determined by the annealing schedule that depends on the solution *u pointwise*

Bayesian criterion - Adaptive Regularization

$$u^* = \arg\max_u \ell(u)^{\lambda(u)}\, q^{(1-\lambda(u))}(u) \tag{22}$$

$$= \arg\min_u \int_\Omega e^{-\frac{\rho(u)}{\beta}} \rho(u)\,\mathrm{d}x + \int_\Omega \left(1 - e^{-\frac{\rho(u)}{\beta}}\right)\gamma(u)\,\mathrm{d}x, \tag{23}$$

- $\ell(u(x)) = e^{-\rho(u(x))}$
- $q(u(x)) = e^{-\gamma(u(x))}$
- $\lambda(u(x)) = e^{-\frac{\rho(u(x))}{\beta}}$ where $\beta > 0$ is a control parameter for the variance of $\rho(u)$

Energy Functional with Adaptive Regularization

$$\mathcal{E}(u;\beta) = \int_\Omega \lambda(u(x))\,\rho(u(x)) + (1-\lambda(u(x)))\,\gamma(u(x))\,\mathrm{d}x \quad (24)$$

$$\lambda(u(x)) = \exp\left(-\frac{\rho(u(x))}{\beta}\right) \quad (25)$$

$\rho(u(x))$ is large

large residual

$\lambda(u(x))$ is small

more rely on the regularization

$\rho(u(x))$ is small

small residual

$\lambda(u(x))$ is large

more rely on the data fidelity

Sparsity Constraint on Weighting Function λ

$$\nu(x) = \exp\left(-\frac{\rho(u(x))}{\beta}\right) \tag{26}$$

$$\lambda(x) = \arg\min_{\lambda} \frac{1}{2}\|\nu(x) - \lambda\|_2^2 + \alpha\|\lambda\|_1 \tag{27}$$

- $\beta > 0$ is a control parameter related to the variation of the residual $\rho(u)$
- $0 < \alpha < 1$ is a constant parameter to control the degree of sparsity in the weighting function λ
- λ is obtained by the solution of the Lasso problem

Application to Imaging Problems

Huber-Huber model

Energy Functional

$$\mathcal{E}(u) = \int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x + \int_\Omega (1-\lambda(x))\,\gamma(u(x))\,\mathrm{d}x \tag{28}$$

$$\rho(u(x)) = \phi_\mu(u(x); f(x)) \tag{29}$$

$$\gamma(u(x)) = \phi_\eta(\nabla u(x)) \tag{30}$$

Huber function

$$\phi_\mu(x) = \begin{cases} \frac{1}{2\mu}x^2 & : |x| \le \mu, \\ |x| - \frac{\mu}{2} & : |x| > \mu \end{cases} \tag{31}$$

where μ is a threshold parameter

Image Segmentation

Data Fidelity

$$\mathcal{D}_{\lambda_i}(u_i, c_i) = \int_\Omega \lambda_i(x)\, \rho(u_i(x), c_i)\, \mathrm{d}x \tag{32}$$

$$\rho(u_i(x), c_i) = \phi_\mu(f(x) - c_i)\, u_i(x) \tag{33}$$

Regularization

$$\mathcal{R}_{\lambda_i}(u_i) = \int_\Omega (1 - \lambda_i(x))\, \gamma(u_i(x))\, \mathrm{d}x \tag{34}$$

$$\gamma(u_i(x)) = \phi_\eta(\nabla u_i(x)) \tag{35}$$

Weighting Function

$$\lambda_i(x) = \exp\left(-\frac{\rho(u_i(x), c_i)}{\beta}\right) : \textit{Lasso} \tag{36}$$

Optical Flow

Data Fidelity

$$\mathcal{D}_\lambda(u) = \int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x \tag{37}$$

$$\rho(u) = \phi_\mu(f_t(x) - \nabla f_1(x + u_0(x)) \cdot (u(x) - u_0(x))) \tag{38}$$

Regularization

$$\mathcal{R}_\lambda(u) = \int_\Omega (1 - \lambda(x))\,\gamma(u(x))\,\mathrm{d}x \tag{39}$$

$$\gamma(u(x)) = \phi_\eta(\nabla u_1(x)) + \phi_\eta(\nabla u_2(x)) \tag{40}$$

Weighting Function

$$\lambda(x) = \exp\left(-\frac{\rho(u(x))}{\beta}\right) : \textit{Lasso} \tag{41}$$

Denoising

Data Fidelity

$$\mathcal{D}_\lambda(u) = \int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x \tag{42}$$

$$\rho(u(x)) = \phi_\mu(f(x) - u(x)) \tag{43}$$

Regularization

$$\mathcal{R}_\lambda(u) = \int_\Omega (1 - \lambda(x))\,\gamma(u(x))\,\mathrm{d}x \tag{44}$$

$$\gamma(u(x)) = \phi_\eta(\nabla u(x)) \tag{45}$$

Weighting Function

$$\lambda(x) = \exp\left(-\frac{\rho(u(x))}{\beta}\right) : \textit{Lasso} \tag{46}$$

Energy Optimization

Alternating Direction method of Multipliers (ADMM)

$$\int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x + \int_\Omega (1-\lambda(x))\,\gamma(\boxed{u(x)})\,\mathrm{d}x$$

$$\int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x + \int_\Omega (1-\lambda(x))\,\gamma(\boxed{v(x)})\,\mathrm{d}x, \text{ subject to } \boxed{u=v},$$

$$\int_\Omega \lambda(x)\,\rho(u(x))\,\mathrm{d}x + \int_\Omega (1-\lambda(x))\,\gamma(\boxed{v(x)})\,\mathrm{d}x + \boxed{\frac{\theta}{2}\|u-v+w\|_2^2}$$

- We initially apply the variable splitting with a new variable v such that $u = v$
- We add a quadratic constraint leading to the unconstrained augmented Lagrangian

Optimization of Huber Function

Moreau-Yosida regularization of a non-smooth function $|\cdot|$

$$\phi_\mu(x) = \inf_r \left\{ |r| + \frac{1}{2\mu}(x-r)^2 \right\} = \mathrm{prox}_{\mu g}(x) \tag{47}$$

- r is an auxiliary variable to be minimized
- $\mathrm{prox}_{\mu g}(x)$ is the proximal operator associated with a convex function $g(x) = \|x\|_1$
- The solution of the above proximal operator $\mathrm{prox}_{\mu g}(x)$ can be obtained by the soft shrinkage operator $\mathcal{T}(x|\mu)$:

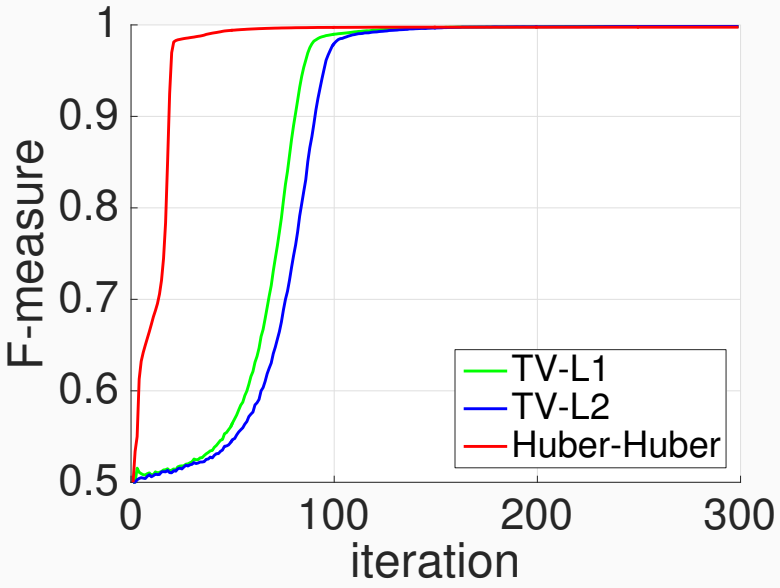
$$\mathcal{T}(x\,|\,\mu) = \begin{cases} x-\mu & : x > \mu \\ 0 & : \|x\|_1 \le \mu \\ x+\mu & : x < -\mu \end{cases} \tag{48}$$

Optimization Algorithm via ADMM

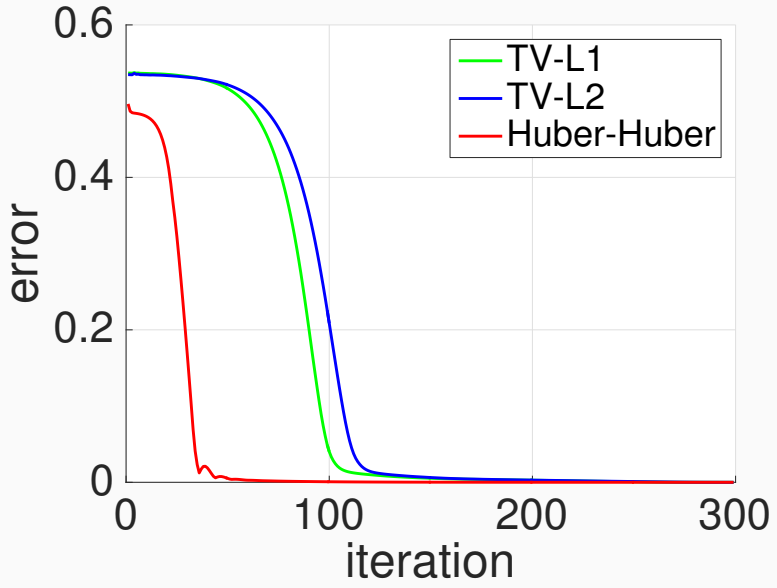
$$
\begin{aligned}
r^{k+1} &:= \operatorname*{argmin}_{r} \rho(u^k, r) \\
z^{k+1} &:= \operatorname*{argmin}_{z} \gamma(v^k, z) \\
u^{k+1} &:= \operatorname*{argmin}_{u} \int_{\Omega} \lambda^{k+1}\, \rho(u, r^{k+1})\, \mathrm{d}x + \frac{\theta}{2} \| u - v^{k+1} + w^k \|_2^2 \\
v^{k+1} &:= \operatorname*{argmin}_{v} \int_{\Omega} (1 - \lambda^{k+1})\, \gamma(v, z^{k+1})\, \mathrm{d}x + \frac{\theta}{2} \| u^k - v + w^k \|_2^2 \\
w^{k+1} &:= w^k + u^{k+1} - v^{k+1} \\
\nu^{k+1} &:= \exp\left(-\frac{\rho(u^{k+1}, r^{k+1})}{\beta} \right) \\
\lambda^{k+1} &:= \operatorname*{argmin}_{\lambda} \frac{1}{2} \| \nu^{k+1} - \lambda \|_2^2 + \alpha \| \lambda \|_1
\end{aligned}
$$

Experimental Results

Comparison of Image Model - Bipartitioning



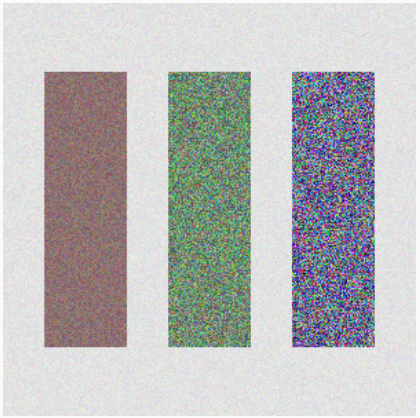
(a) accuracy



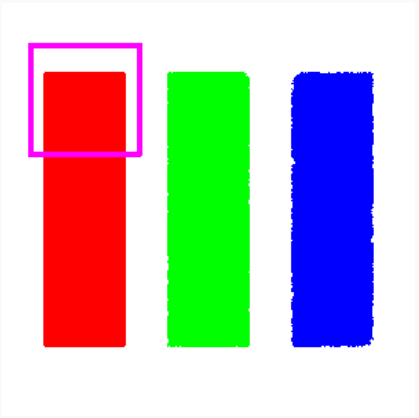
(b) error

- Images are taken from Berkeley segmentation dataset
- Bi-partitioning segmentation is performed based on TV-L_1 (green), TV-L_2 (blue), our H^2 (red) models
- F-measure (left) and error (right) are computed over iteration

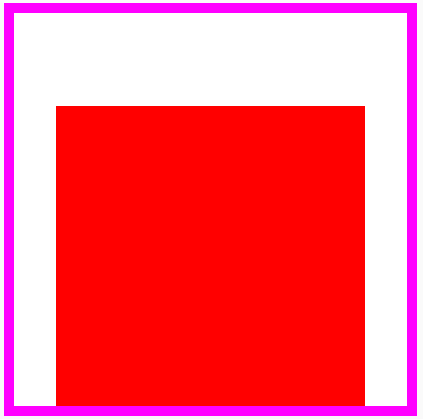
Adaptive Regularization - Segmentation



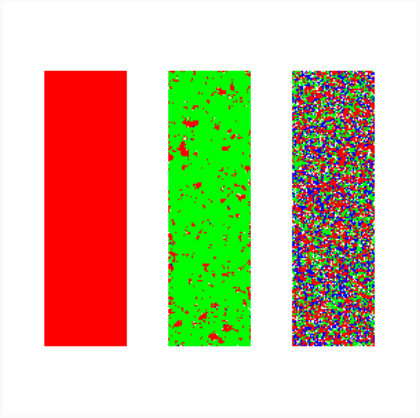
(a) Input



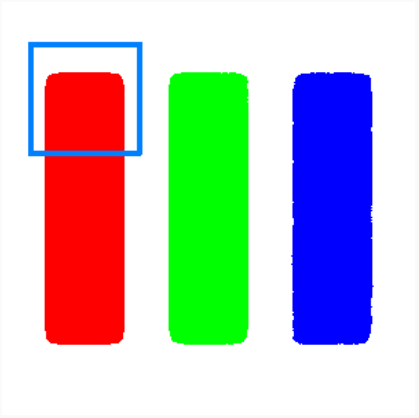
(b) Ours (adaptive)



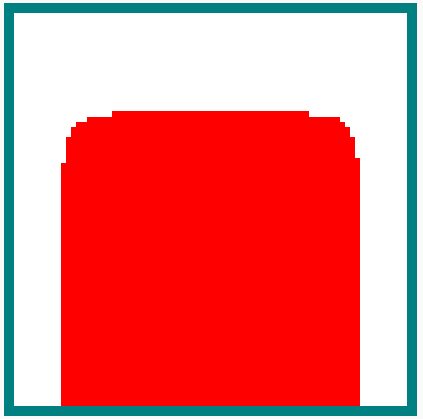
(c) Zoom in of (b)



(d) Small (global)



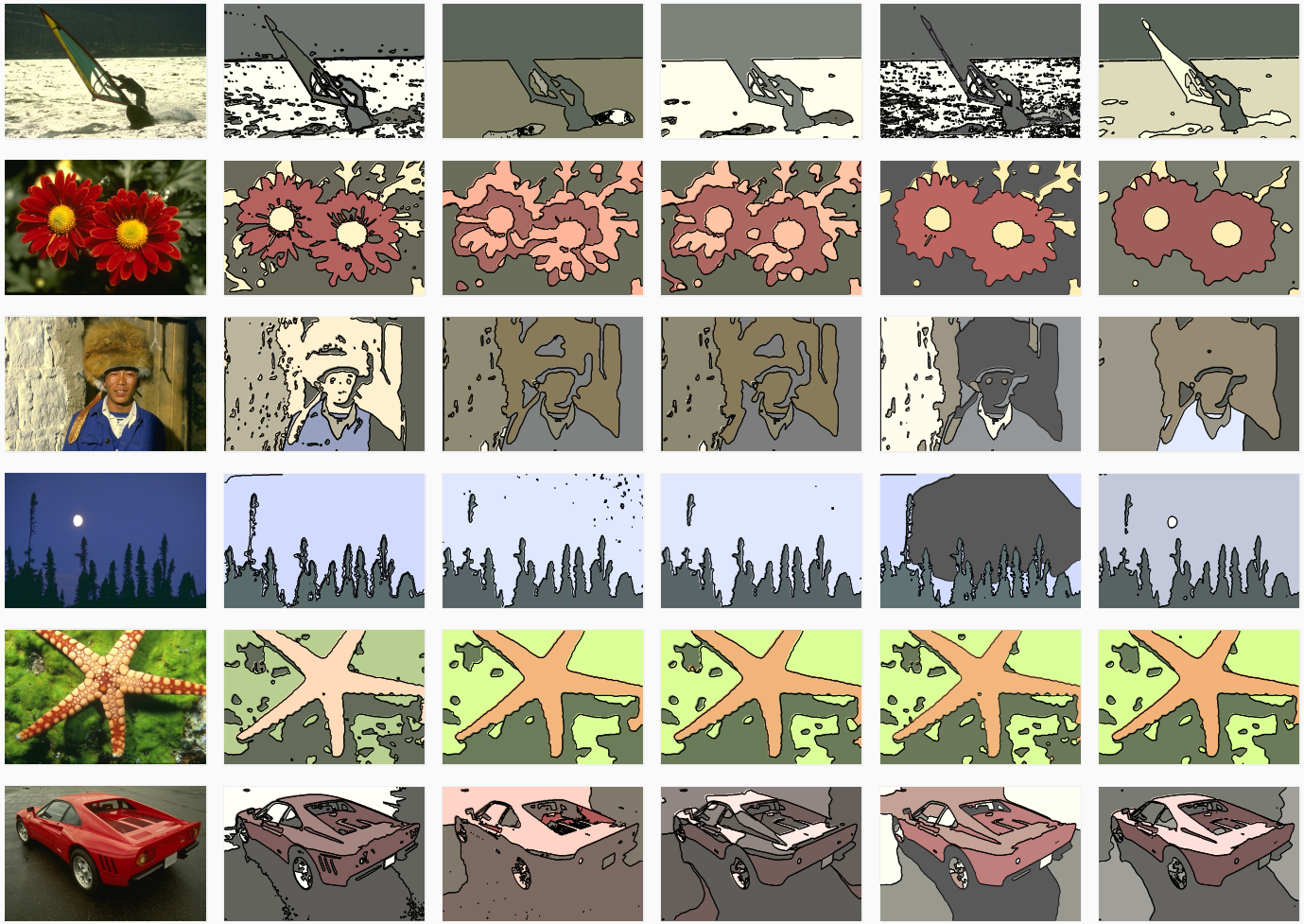
(e) Large (global)



(f) Zoom in of (e)

- Illustrative comparison of constant and adaptive regularization

Experimental Results - Segmentation



(a) Input (b) FL[1] (c) TV[2] (d) VTV[3] (e) PC[4] (f) Ours

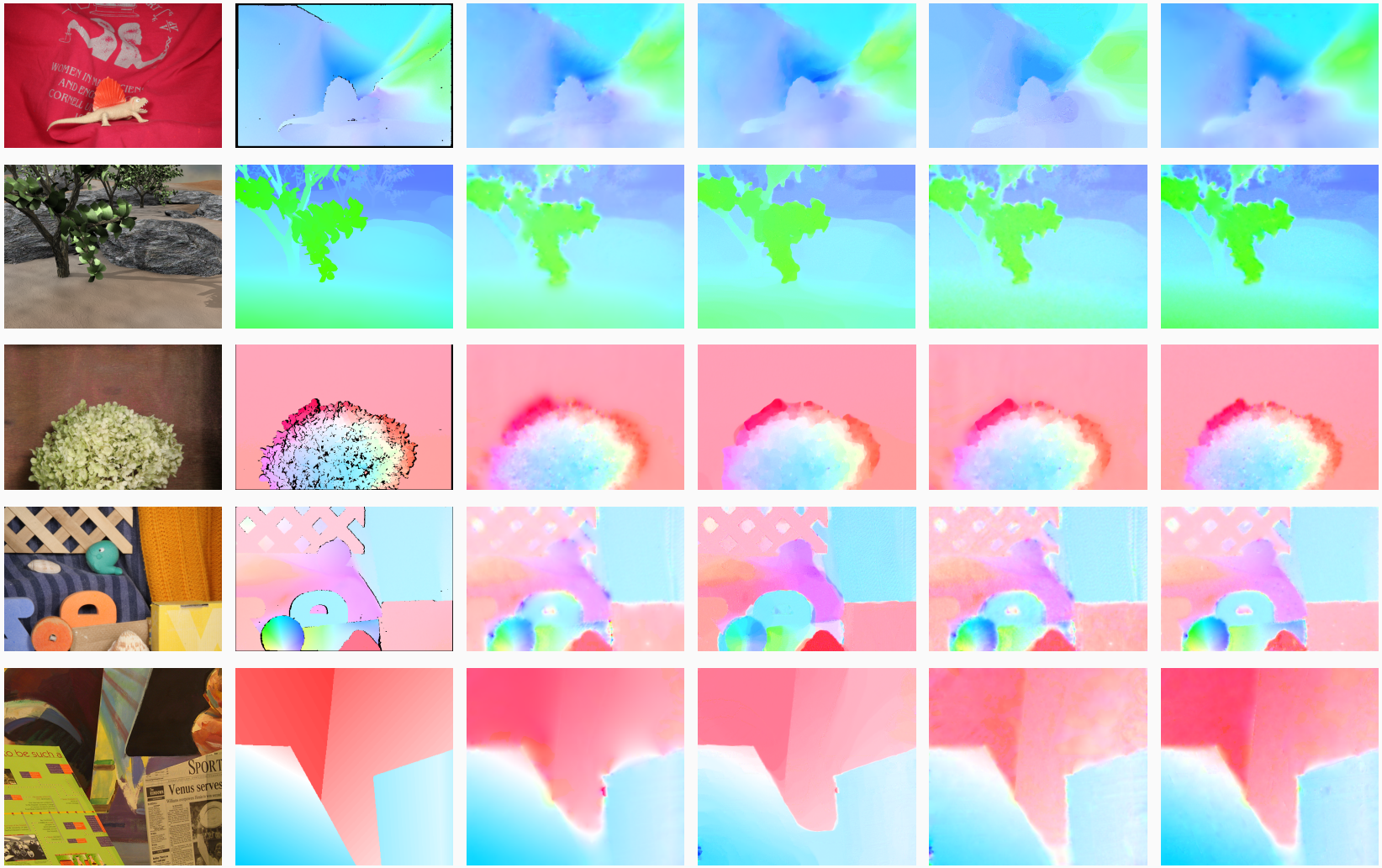
[1] **sundaramoorthi2014fast**.

[2] **zach2008fast**.

[3] **lellmann:continuous:siam:2011**.

[4] **chambolle2012convex**.

Quantitative Evaluation (Average Angular Error) - Motion



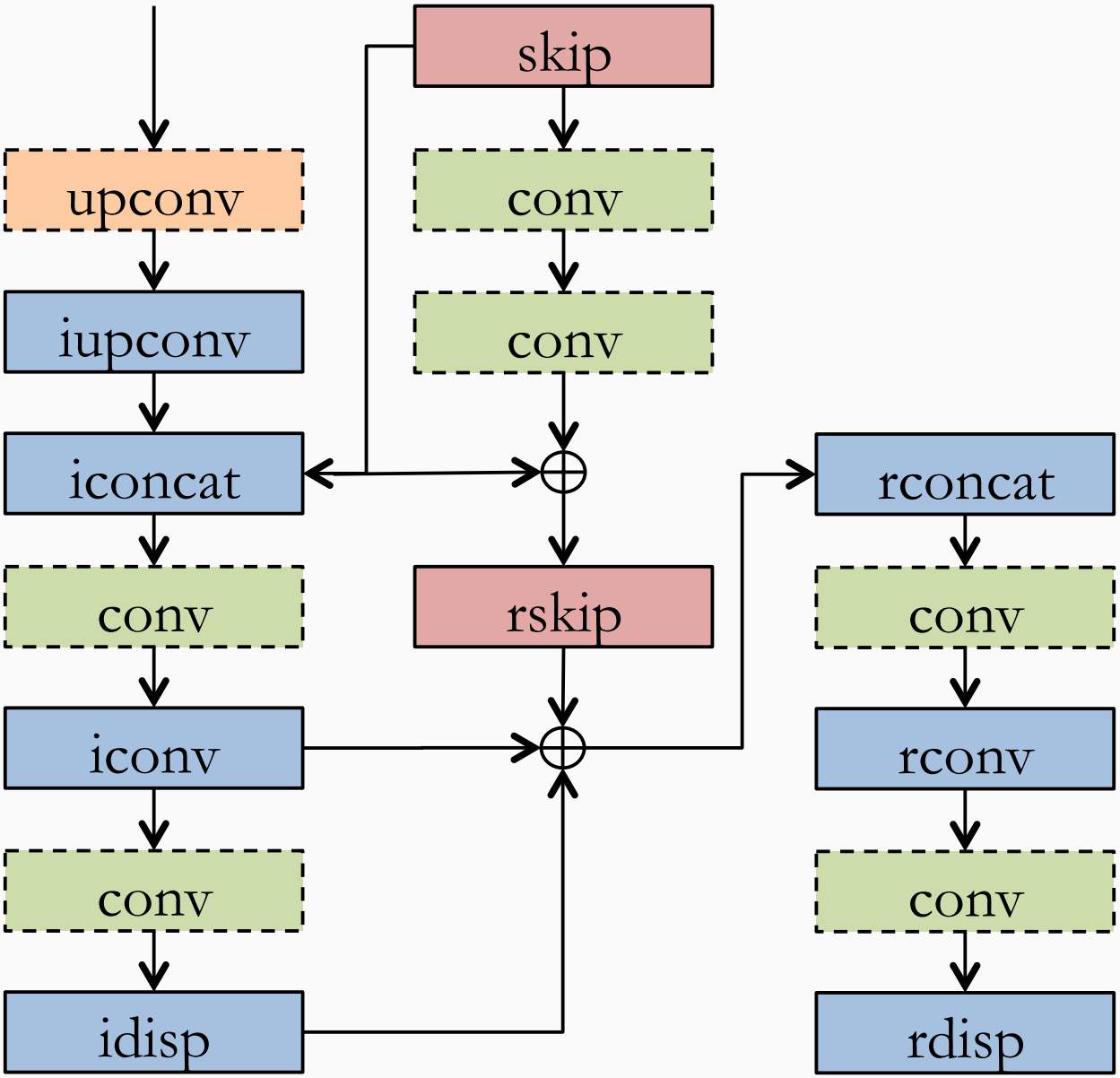
(a) Input (b) GT (c) HS[5] (d) TV[6] (e) HTV[7] (f) Ours

[5] **horn1981determining**.

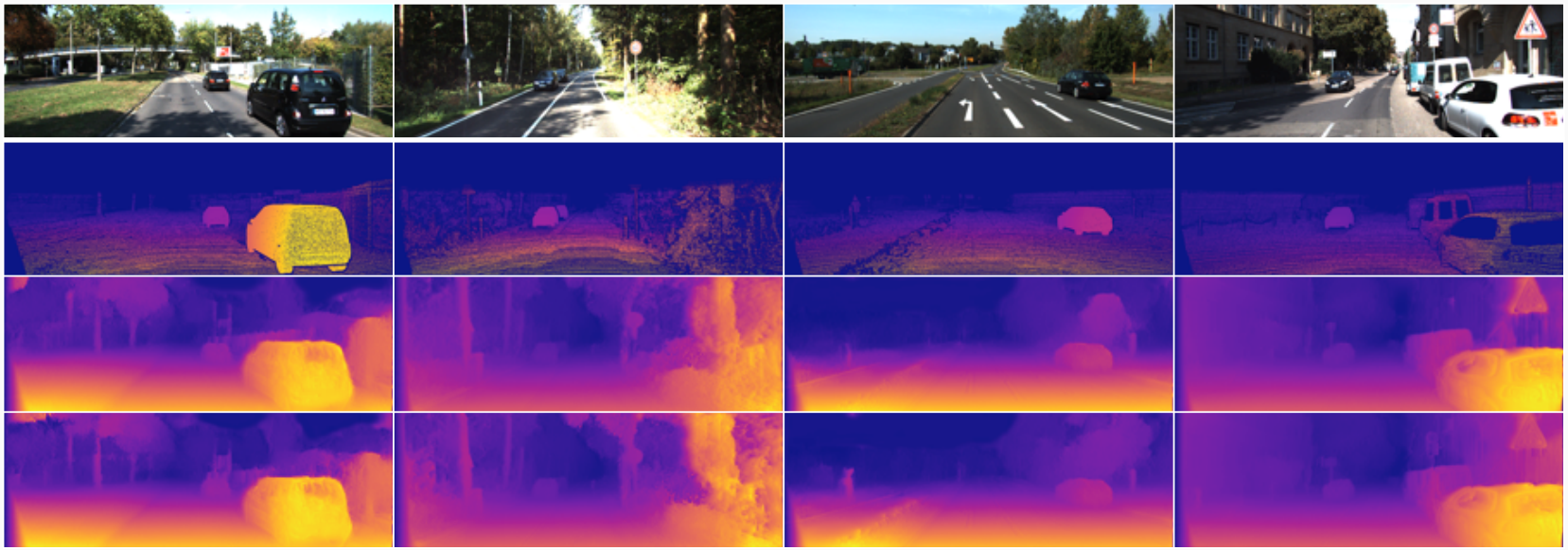
[6] **zach2007duality**.

[7] **werlberger2009anisotropic**.

Deep Learning Application - Depth Prediction



Deep Learning Application - Depth Prediction (KITTI)



Adaptive Regularization in Deep Learning framework

- Training data : $\{(x_i, y_i)\}_{i=1}^n, \quad x_i \in X, y_i \in Y$
- Prediction function : $h_w : X \to Y$
- Model parameters : $w = (w_1, w_2, \cdots, w_m) \in \mathbb{R}^m$

$$\mathcal{L}(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) + \lambda \gamma(w) \tag{49}$$

- $f_i(w)$ is a data fidelity incurred by a set of model parameters w for a sample pair (x_i, y_i)
- $f_i(w)$ measures discrepancy between $h_w(x_i)$ and y_i
- $\gamma(w)$ is a regularization on model parameters w

Adaptive Regularization in Deep Learning framework

Static Regularization (Weight Decay)

$$\mathcal{L}(w) = \rho(w) + \frac{\lambda}{2} \|w\|_2^2, \tag{50}$$

Adaptive Regularization

$$\mathcal{L}(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) + \frac{\lambda}{2} \|\theta \odot w\|_2^2, \tag{51}$$

- $\theta = (\theta_1, \theta_2, \cdots, \theta_m) \in \mathbb{R}^m$
- $\odot$ denotes the Hadamard product
- $\theta \odot w = (\theta_1 w_1, \theta_2 w_2, \cdots, \theta_m w_m)$

Adaptive Weight Decay

Weight Decay based on Residual

$$\tilde{g}_j^t = \frac{|g_j^t| - \mu_l^t}{\sigma_l^t}, \tag{52}$$

$$\theta_j^t \propto \tilde{g}_j^t, \tag{53}$$

- g_j^t is the gradient of the data fidelity with respect to the parameter w_j at iteration t
- l denotes the index of the layer that includes the parameter w_j
- μ_l^t and σ_l^t denotes the mean and standard deviation of all the gradient norms for the parameters within the layer l at iteration t

Adaptive Weight Decay

Adaptive Decay Rate

$$\theta_j^t = S(\tilde{g}_j^t; \alpha) = \frac{2}{1 + \exp(-\alpha\, \tilde{g}_j^t)}, \tag{54}$$

- S is a scaled signmoid function
- α determines the slope of the decay rate
- $\tilde{g}_j^t$ is normalized to have mean 0 and standard deviation 1.
- θ_j^t ranges from 0 to 2

Numerical Results - Validation Accuracy for MNIST

(1) Validation accuracy for MNIST

	NN-2						NN-3					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	98.53	98.22	98.19	98.16	98.15	**98.56**	98.66	98.31	98.26	98.29	98.23	**98.69**
max	98.63	98.36	98.31	98.31	98.35	**98.72**	98.80	98.49	98.47	98.41	98.45	**98.82**
	LeNet-4						VGG-9					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	99.31	99.30	99.27	99.23	99.17	**99.32**	99.62	99.37	99.37	99.58	99.52	**99.63**
max	**99.48**	99.39	99.37	99.38	99.28	99.45	**99.71**	99.50	99.43	99.63	99.59	99.70

Numerical Results - Validation Accuracy for Fashion-MNIST

(2) Validation accuracy for Fashion-MNIST

	NN-2						NN-3					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	89.23	88.89	88.98	87.63	89.12	**89.49**	89.71	89.17	89.26	88.13	89.24	**89.95**
max	89.50	89.15	89.28	87.83	89.44	**89.84**	89.87	89.54	89.49	88.32	89.56	**90.23**
	LeNet-4						VGG-9					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	90.65	90.81	90.78	89.76	90.12	**90.87**	93.45	91.97	92.08	93.33	93.05	**93.51**
max	91.36	91.30	91.23	90.54	90.48	**91.51**	93.73	92.36	92.46	93.72	93.37	**93.83**

Numerical Results - Validation Accuracy for CIFAR-10

(3) Validation accuracy for CIFAR-10

	ResNet-18						ResNet-50					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	94.70	90.72	90.92	91.46	93.07	**94.80**	94.61	91.22	91.26	90.76	92.82	**94.71**
max	94.98	91.25	91.36	91.99	93.38	**95.04**	95.16	91.82	91.73	91.38	93.36	**95.22**
	GoogLeNet						DenseConv					
	SGD	RMS	Adam	eSGD	aSGD	Ours	SGD	RMS	Adam	eSGD	aSGD	Ours
ave	94.91	90.48	90.62	93.00	93.39	**95.17**	94.72	83.17	83.80	88.05	90.27	**94.91**
max	95.43	90.94	90.97	93.45	93.79	**95.50**	95.08	83.60	84.40	88.60	90.56	**95.22**

Summary

Summary

- We have introduced an adaptive regularization scheme based on the current local data fit to the model during the iterative optimization
- We have presented classical imaging problems based on the Huber-Huber model
- We have presented an efficient optimization algorithm based on ADMM framework
- Numerical experiments have demonstrated the effectiveness and robustness of the adaptive regularization scheme