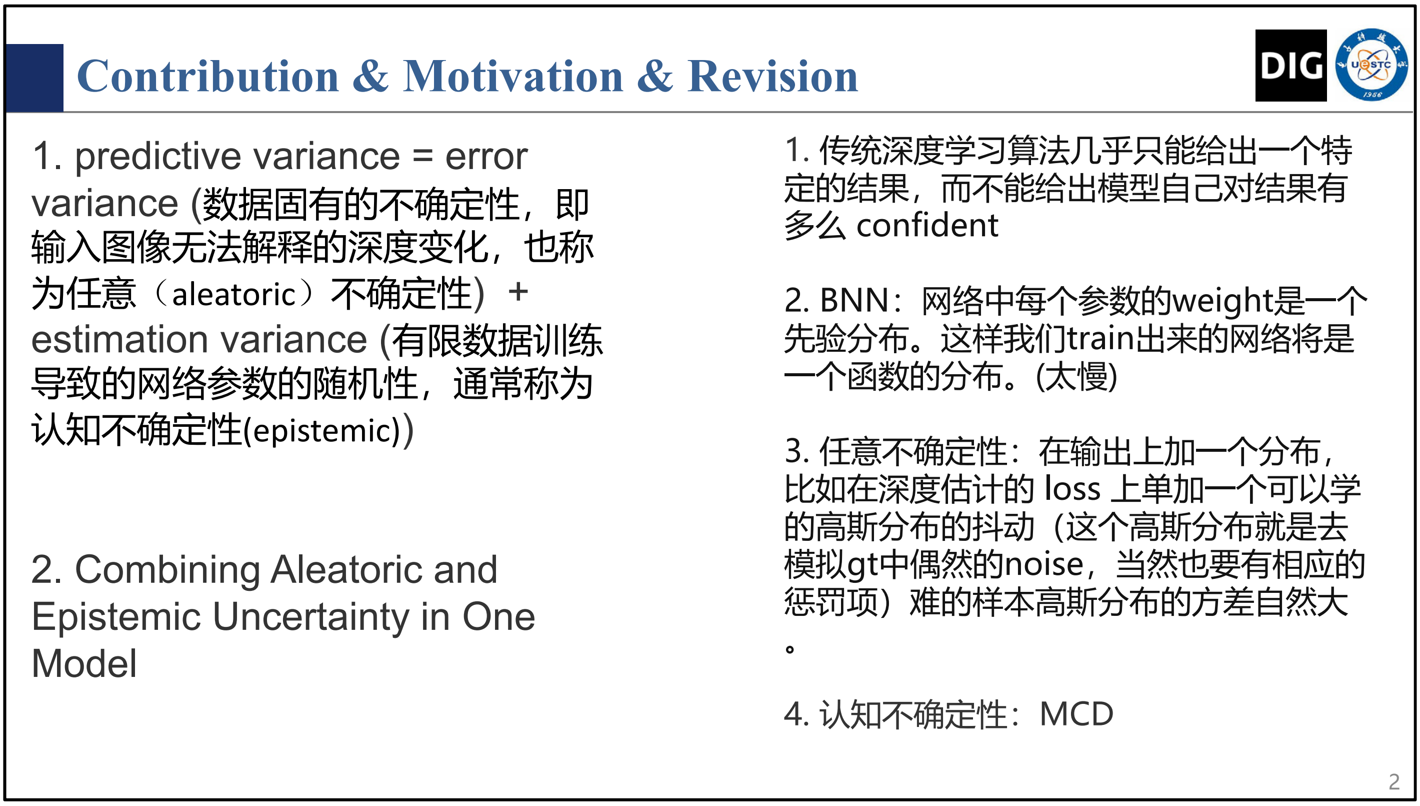
What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? (NIPS 2017 4000+ cites)

2023.12.07



Contribution & Motivation & Revision

1. predictive variance = error variance (数据固有的不确定性，即输入图像无法解释的深度变化，也称为任意（aleatoric）不确定性） + estimation variance (有限数据训练导致的网络参数的随机性，通常称为认知不确定性(epistemic))

2. Combining Aleatoric and Epistemic Uncertainty in One Model

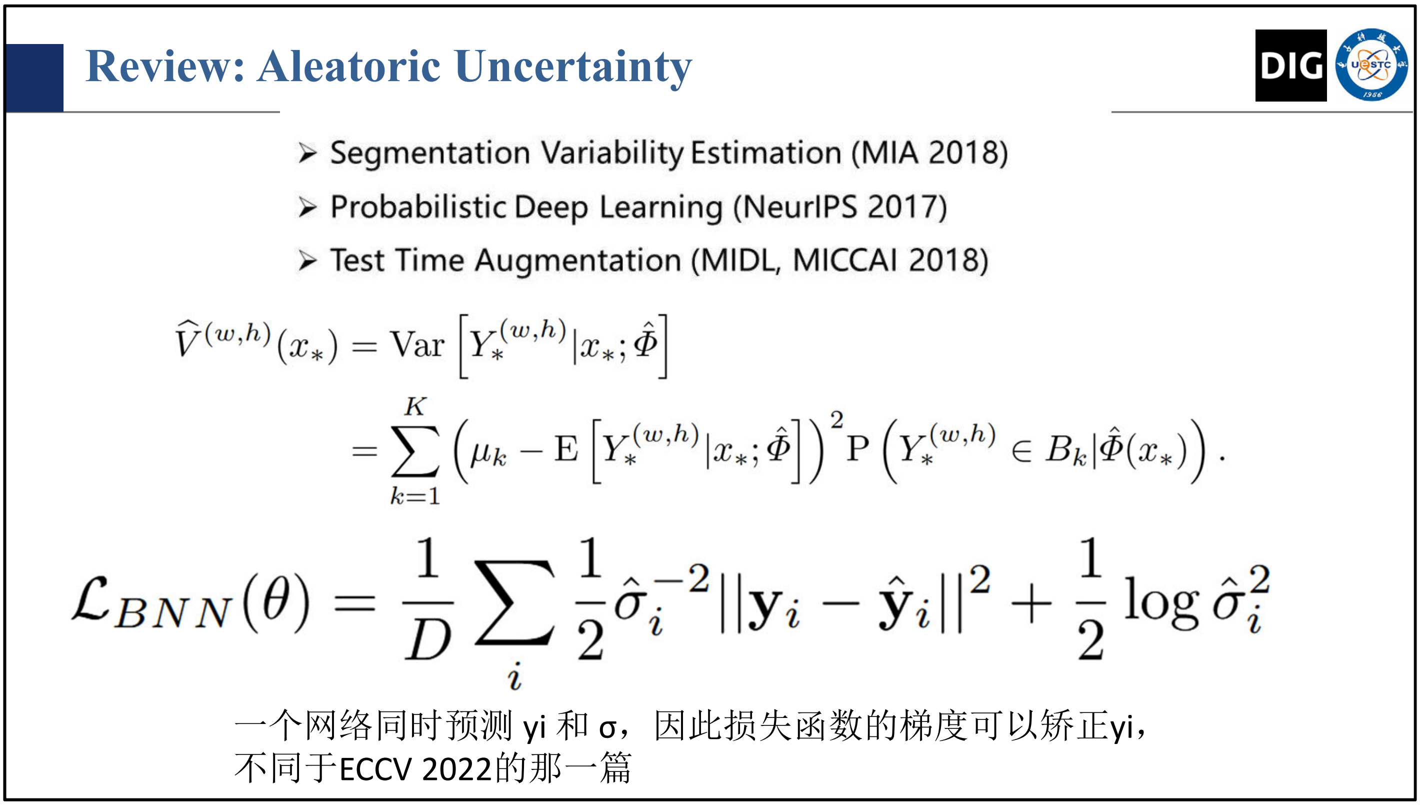
1. 传统深度学习算法几乎只能给出一个特定的结果，而不能给出模型自己对结果有多么 confident

2. BNN：网络中每个参数的weight是一个先验分布。这样我们train出来的网络将是一个函数的分布。(太慢)

3. 任意不确定性：在输出上加一个分布，比如在深度估计的 loss 上单加一个可以学的高斯分布的抖动（这个高斯分布就是去模拟gt中偶然的noise，当然也要有相应的惩罚项）难的样本高斯分布的方差自然大。

4. 认知不确定性：MCD

在输出上加一个分布，比如在深度估计的结果上 加一个（可以学的）高斯分布的抖动（这个高斯分布就是去模拟实际观测中偶然的noise）。难的样本 高斯分布的 方差自然大。



Review: Aleatoric Uncertainty

- ➢ Segmentation Variability Estimation (MIA 2018)
- ➢ Probabilistic Deep Learning (NeurIPS 2017)
- ➢ Test Time Augmentation (MIDL, MICCAI 2018)

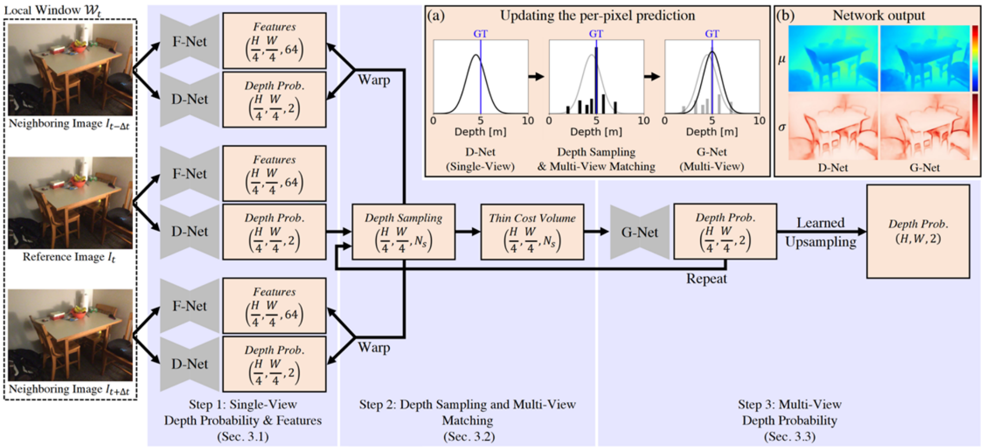
$$\widehat{V}^{(w,h)}(x_*) = \mathrm{Var}\left[Y_*^{(w,h)}|x_*;\hat{\Phi}\right] = \sum_{k=1}^{K}\left(\mu_k - \mathrm{E}\left[Y_*^{(w,h)}|x_*;\hat{\Phi}\right]\right)^2 \mathrm{P}\left(Y_*^{(w,h)} \in B_k|\hat{\Phi}(x_*)\right).$$

$$\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_i \frac{1}{2}\hat{\sigma}_i^{-2}||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \frac{1}{2}\log\hat{\sigma}_i^2$$

一个网络同时预测 yi 和 σ，因此损失函数的梯度可以矫正yi，
不同于ECCV 2022的那一篇

这个loss的推导是由最小负对数似然函数而来的，起初EU中的方差，是假定数据中含有固定的噪声，后续考虑到模型可以自动学习这个AU，就将其作为其中一份子进行学习出来。公式是基于满足高斯分布推导而来的。

Review: Aleatoric Uncertainty


Local Window $\mathcal{W}_t$
Neighboring Image $I_{t-\Delta t}$
Reference Image I_t
Neighboring Image $I_{t+\Delta t}$
F-Net
D-Net
Features
$\left(\frac{H}{4},\frac{W}{4},64\right)$
Depth Prob.
$\left(\frac{H}{4},\frac{W}{4},2\right)$
Warp
Depth Sampling
$\left(\frac{H}{4},\frac{W}{4},N_s\right)$
Thin Cost Volume
$\left(\frac{H}{4},\frac{W}{4},N_s\right)$
G-Net
Learned
Upsampling
Depth Prob.
$(H,W,2)$
Repeat
(a) Updating the per-pixel prediction
GT
0
5
10
Depth [m]
D-Net
(Single-View)
Depth Sampling
& Multi-View Matching
G-Net
(Multi-View)
(b) Network output
μ
σ
D-Net
G-Net
Step 1: Single-View
Depth Probability & Features
(Sec. 3.1)
Step 2: Depth Sampling and Multi-View
Matching
(Sec. 3.2)
Step 3: Multi-View
Depth Probability
(Sec. 3.3)

$$L_{u,v}(d^{\text{gt}}_{u,v}|I_t) = \frac{1}{2}\log\sigma^2_{u,v}(I_t) + \frac{\left(d^{\text{gt}}_{u,v} - \mu_{u,v}(I_t)\right)^2}{2\sigma^2_{u,v}(I_t)}.$$

这个loss的推导是由最小负对数似然函数而来的，起初EU中的方差，是假定数据中含有固定的噪声，后续考虑到模型可以自动学习这个AU，就将其作为其中一份子进行学习出来。公式是基于满足高斯分布推导而来的。

Review: Aleatoric Uncertainty

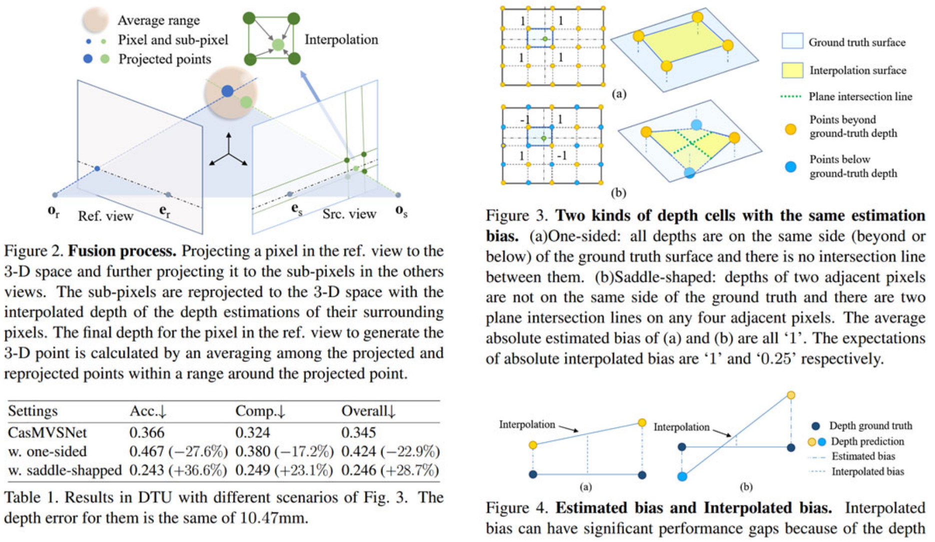


Figure 2. **Fusion process.** Projecting a pixel in the ref. view to the 3-D space and further projecting it to the sub-pixels in the others views. The sub-pixels are reprojected to the 3-D space with the interpolated depth of the depth estimations of their surrounding pixels. The final depth for the pixel in the ref. view to generate the 3-D point is calculated by an averaging among the projected and reprojected points within a range around the projected point.

Settings	Acc.↓	Comp.↓	Overall↓
CasMVSNet	0.366	0.324	0.345
w. one-sided	0.467 (−27.6%)	0.380 (−17.2%)	0.424 (−22.9%)
w. saddle-shapped	0.243 (+36.6%)	0.249 (+23.1%)	0.246 (+28.7%)

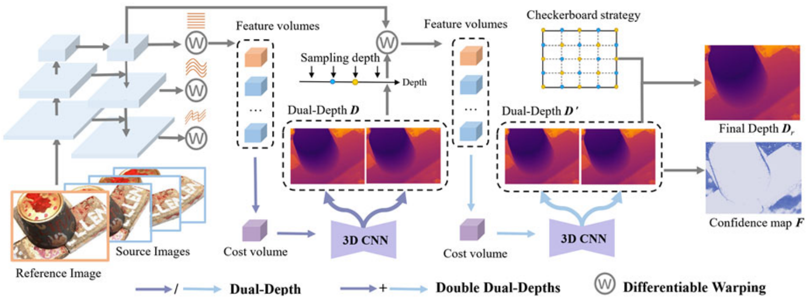
Table 1. Results in DTU with different scenarios of Fig. 3. The depth error for them is the same of 10.47mm.

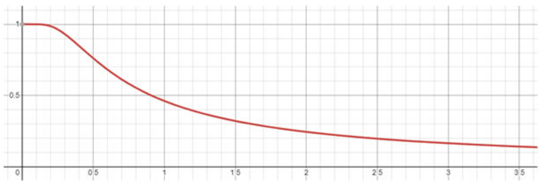
Figure 3. **Two kinds of depth cells with the same estimation bias.** (a)One-sided: all depths are on the same side (beyond or below) of the ground truth surface and there is no intersection line between them. (b)Saddle-shaped: depths of two adjacent pixels are not on the same side of the ground truth and there are two plane intersection lines on any four adjacent pixels. The average absolute estimated bias of (a) and (b) are all ‘1’. The expectations of absolute interpolated bias are ‘1’ and ‘0.25’ respectively.

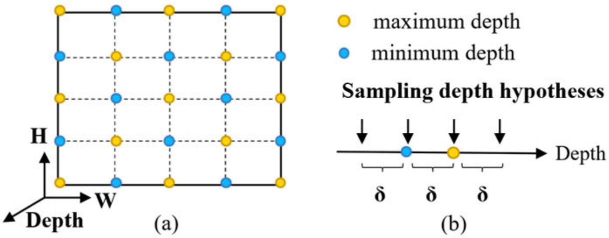
Figure 4. **Estimated bias and Interpolated bias.** Interpolated bias can have significant performance gaps because of the depth

损失函数要求 interval 和 最大的estimation bias 差距不大，如果和gt相比的estimation bias变大，第一项也需要变大，因此保证在两侧。

Review: Aleatoric Uncertainty







$$
\begin{aligned}
L_{int}(\boldsymbol{D}, \boldsymbol{D}_g) = L_1(&|\mathtt{max}(\boldsymbol{D}) - \mathtt{min}(\boldsymbol{D})|, \\
&max(|\mathtt{max}(\boldsymbol{D}) - \boldsymbol{D}_g|, |\mathtt{min}(\boldsymbol{D}) - \boldsymbol{D}_g|))\,,
\end{aligned}
$$

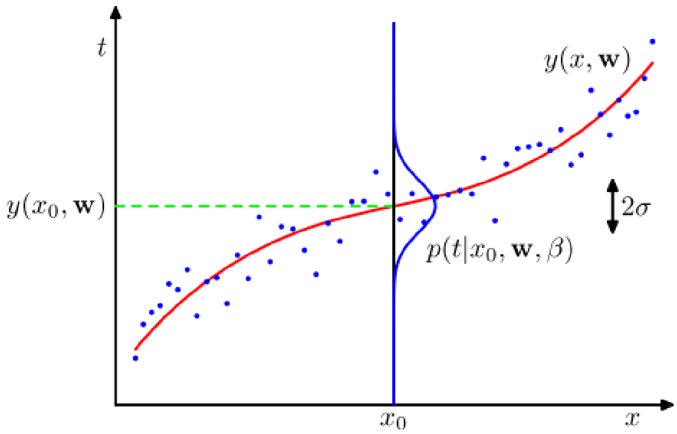
$$
\boldsymbol{F}(x,y) = 2\mathtt{sigmoid}(\frac{1}{\boldsymbol{U}(x,y)}) - 1 \in (0,1)\ ,
$$

where $\boldsymbol{U}(x,y) \;=\; |\mathtt{max}(\boldsymbol{D})(x,y) - \mathtt{min}(\boldsymbol{D})(x,y)|.$

损失函数要求 interval 和 最大的estimation bias 差距不大，如果和gt相比的estimation bias变大，第一项也需要变大，因此保证在两侧。

Review: Gaussian Likelihood

极大似然估计： $p(y|\theta)$,找到一组参数，使得在这套参数下，我们观测到这批数据的概率最大。[3]



Review: Gaussian Likelihood

$$p(y|x,\theta) = \prod_{i=1}^{N} p(y_i|x_i,\theta)$$

$$\theta^{MLE} = \arg\max_{\theta} \sum_{i=1}^{N} \log p(y_i|x_i,\theta)$$

$$\max_{\theta} \log p(y|x,\theta)$$

$$= \max_{\theta} \sum_{i=1}^{N} \log p(y_i|\hat{y}_i(x_i,\theta), \sigma_i^2(x_i,\theta))$$

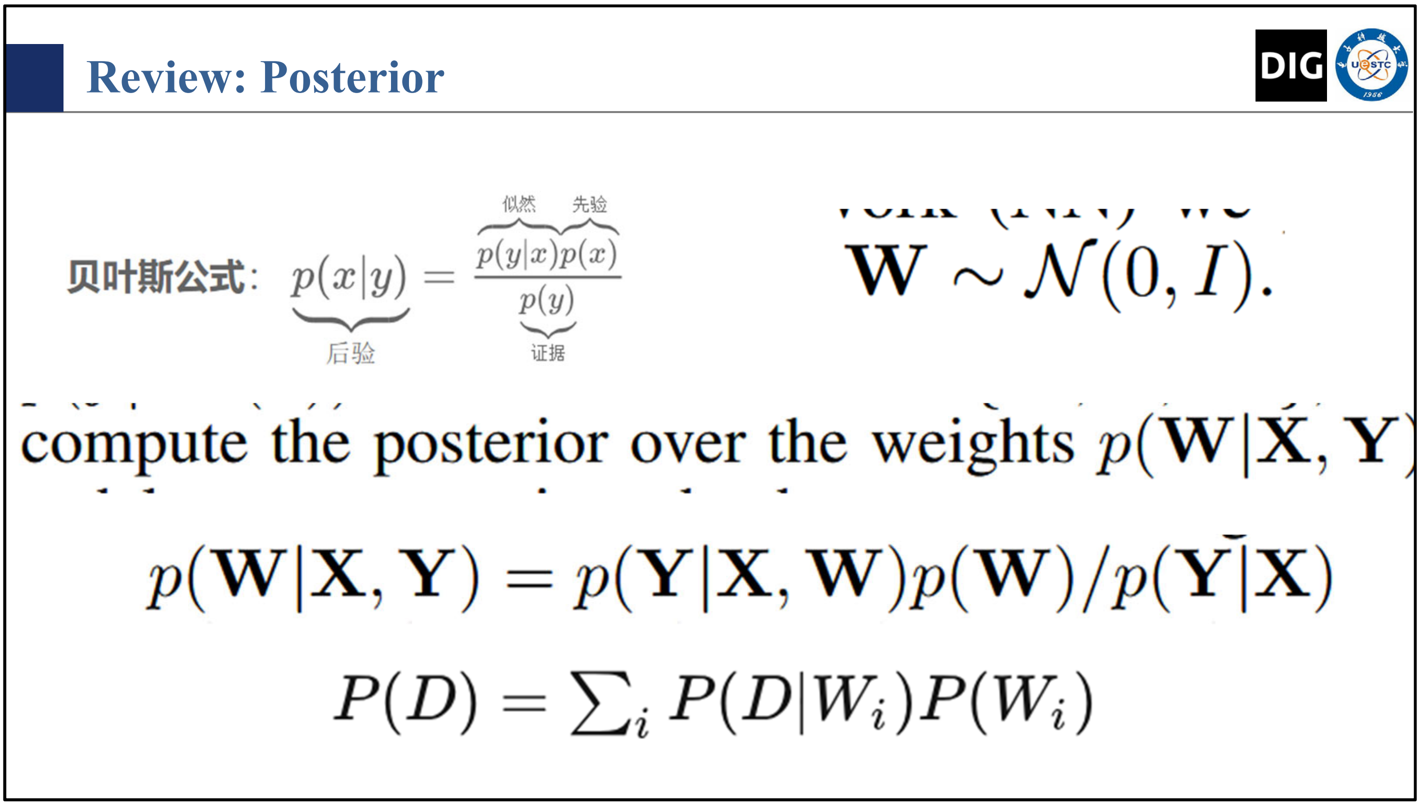
$$= \max_{\theta} \sum_{i=1}^{N} \log \mathcal{N}(\hat{y}_i, \sigma_i^2)$$

$$= \max_{\theta} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{||y_i - \hat{y}_i||^2}{2\sigma_i^2})$$

$$= \max_{\theta} \sum_{i=1}^{N} \{-\frac{||y_i - \hat{y}_i||^2}{2\sigma_i^2} - \frac{\log \sigma_i^2}{2} - \frac{\log 2\pi}{2}\}$$

最大化似然估计相当于最小化损失，即

$$\min_{\theta} \mathcal{L} = \sum_{i=1}^{N} \frac{||y_i - \hat{y}_i||^2}{2\sigma_i^2} + \frac{\log \sigma_i^2}{2}$$



Review: Posterior

贝叶斯公式：$\underbrace{p(x|y)}_{后验} = \frac{\overbrace{p(y|x)}^{似然}\overbrace{p(x)}^{先验}}{\underbrace{p(y)}_{证据}}$

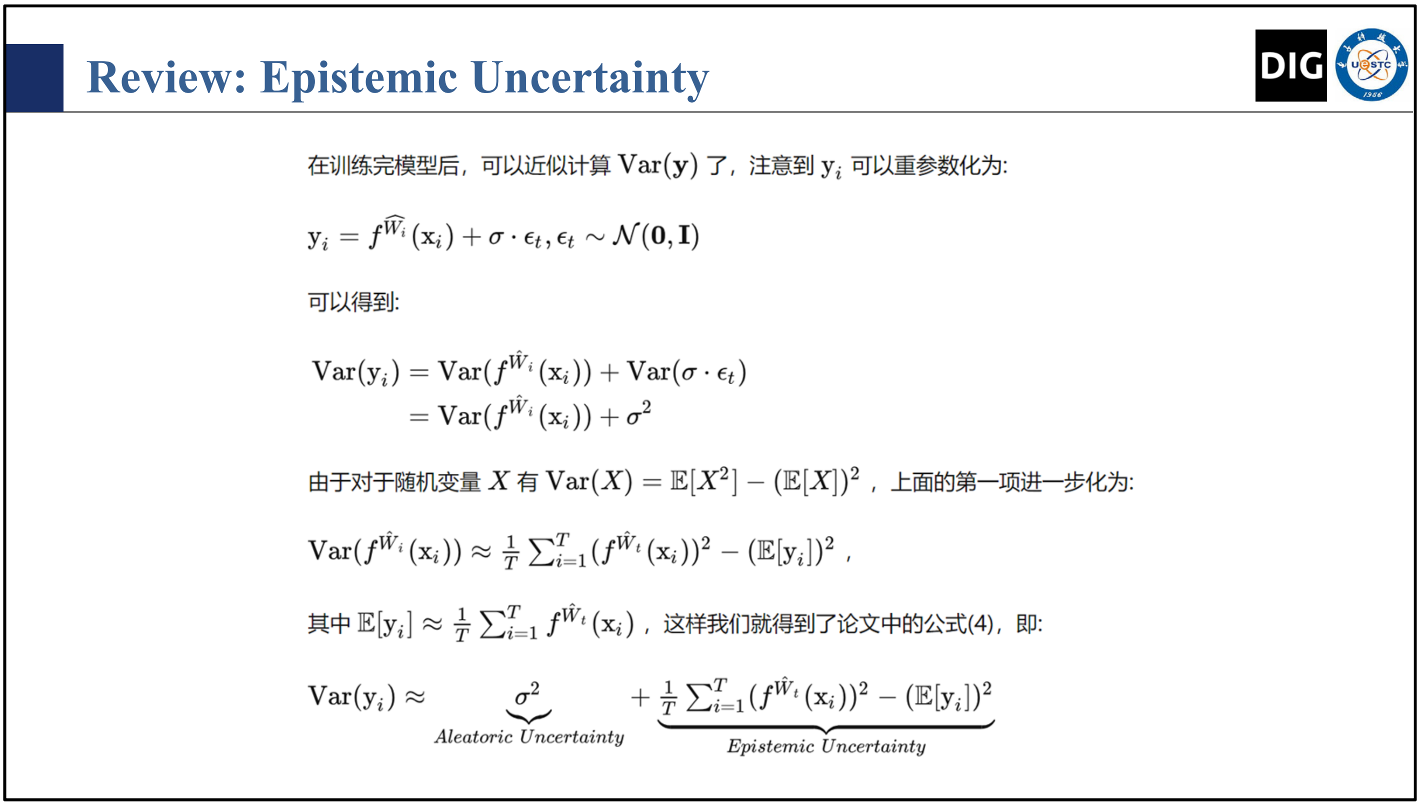
$$\mathbf{W} \sim \mathcal{N}(0, I).$$

compute the posterior over the weights $p(\mathbf{W}|\mathbf{X}, \mathbf{Y})$

$$p(\mathbf{W}|\mathbf{X}, \mathbf{Y}) = p(\mathbf{Y}|\mathbf{X}, \mathbf{W})p(\mathbf{W})/p(\mathbf{Y}|\mathbf{X})$$

$$P(D) = \sum_i P(D|W_i)P(W_i)$$

贝叶斯推理用于计算权重 p(W|X, Y) 的后验
认知不确定性是通过对模型的权重进行先验分布来建模的，然后尝试捕获这些权重在给定某些数据的情况下变化的程度。



Review: Epistemic Uncertainty

在训练完模型后，可以近似计算 $\mathrm{Var}(\mathbf{y})$ 了，注意到 y_i 可以重参数化为:

$$\mathrm{y}_i = f^{\widehat{W_i}}(\mathrm{x}_i) + \sigma \cdot \epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

可以得到:

$$\begin{aligned} \mathrm{Var}(\mathrm{y}_i) &= \mathrm{Var}(f^{\hat{W}_i}(\mathrm{x}_i)) + \mathrm{Var}(\sigma \cdot \epsilon_t) \\ &= \mathrm{Var}(f^{\hat{W}_i}(\mathrm{x}_i)) + \sigma^2 \end{aligned}$$

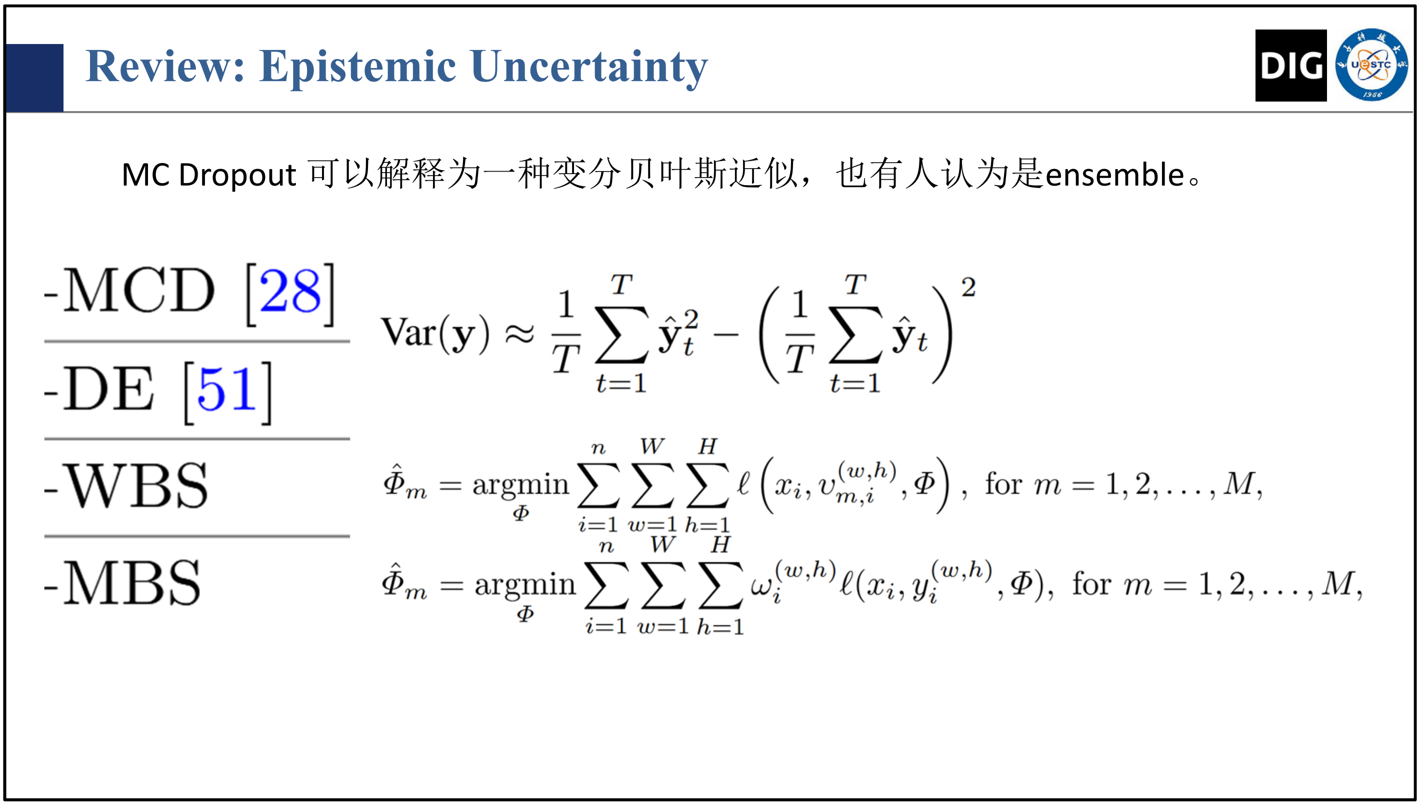
由于对于随机变量 X 有 $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$，上面的第一项进一步化为:

$$\mathrm{Var}(f^{\hat{W}_i}(\mathrm{x}_i)) \approx \frac{1}{T}\sum_{i=1}^{T}(f^{\hat{W}_t}(\mathrm{x}_i))^2 - (\mathbb{E}[\mathrm{y}_i])^2,$$

其中 $\mathbb{E}[\mathrm{y}_i] \approx \frac{1}{T}\sum_{i=1}^{T} f^{\hat{W}_t}(\mathrm{x}_i)$，这样我们就得到了论文中的公式(4)，即:

$$\mathrm{Var}(\mathrm{y}_i) \approx \underbrace{\sigma^2}_{Aleatoric\ Uncertainty} + \underbrace{\frac{1}{T}\sum_{i=1}^{T}(f^{\hat{W}_t}(\mathrm{x}_i))^2 - (\mathbb{E}[\mathrm{y}_i])^2}_{Epistemic\ Uncertainty}$$

直接用交叉熵做分类的话，错误的情况无法区分。
确保远离真实标签的预测比接近真实标签的预测会受到更大的惩罚



Review: Epistemic Uncertainty

MC Dropout 可以解释为一种变分贝叶斯近似，也有人认为是ensemble。

-MCD [28]

-DE [51]

-WBS

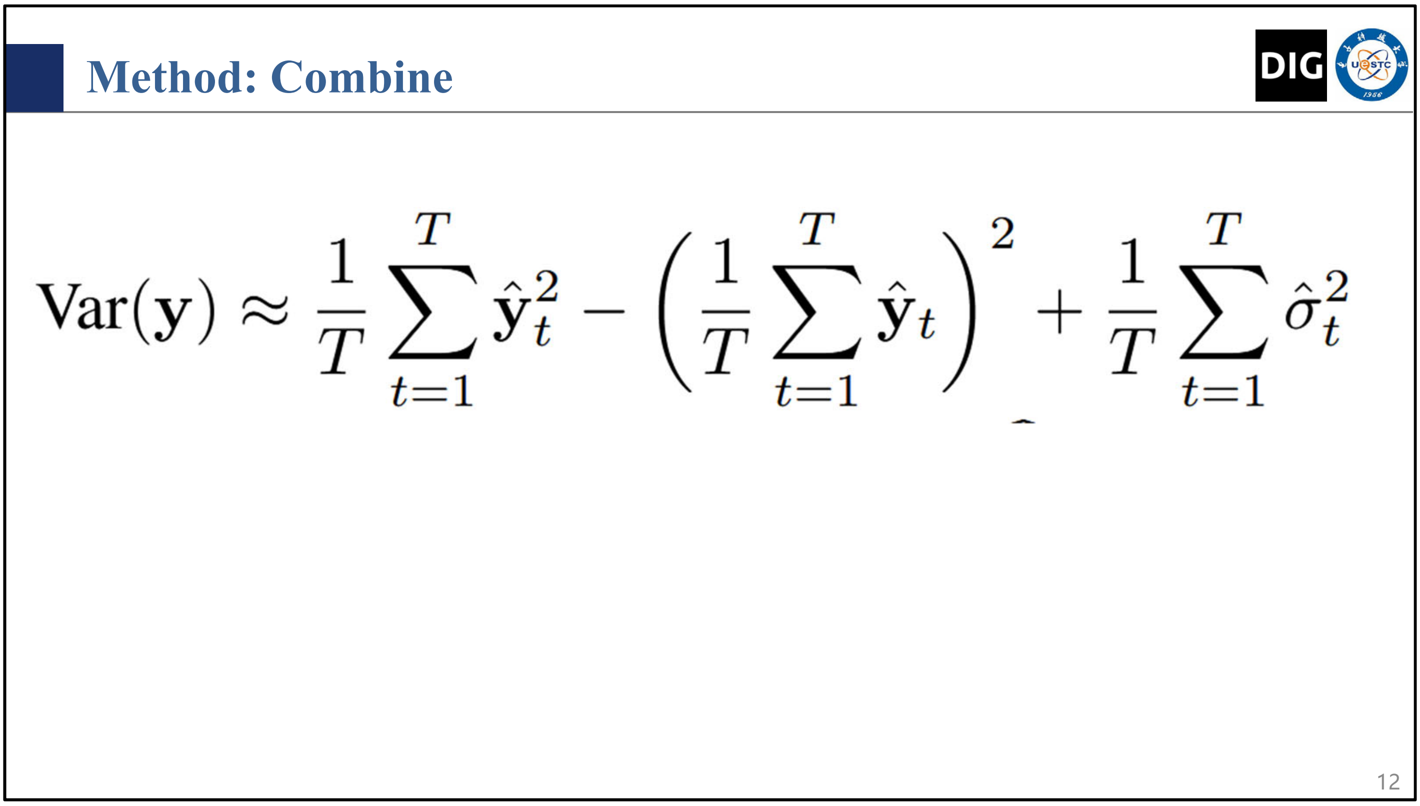
-MBS

$$\mathrm{Var}(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}_t\right)^2$$

$$\hat{\Phi}_m = \underset{\Phi}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{w=1}^{W} \sum_{h=1}^{H} \ell\left(x_i, v_{m,i}^{(w,h)}, \Phi\right), \text{ for } m = 1, 2, \ldots, M,$$

$$\hat{\Phi}_m = \underset{\Phi}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{w=1}^{W} \sum_{h=1}^{H} \omega_i^{(w,h)} \ell(x_i, y_i^{(w,h)}, \Phi), \text{ for } m = 1, 2, \ldots, M,$$

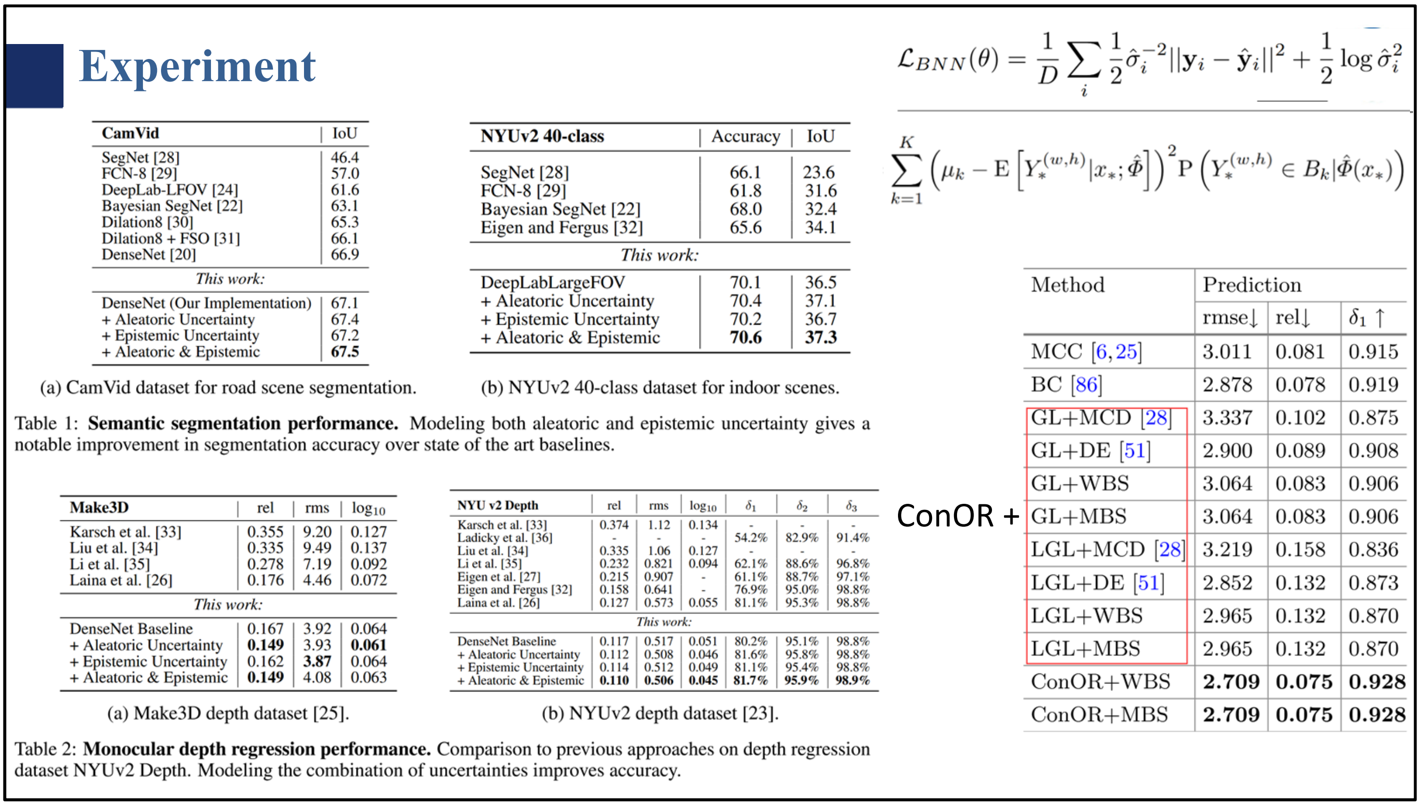
直接用交叉熵做分类的话，错误的情况无法区分。
确保远离真实标签的预测比接近真实标签的预测会受到更大的惩罚



Method: Combine

$$\mathrm{Var}(\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{y}}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_t^2$$

直接用交叉熵做分类的话，错误的情况无法区分。
确保远离真实标签的预测比接近真实标签的预测会受到更大的惩罚



Experiment

$$\mathcal{L}_{BNN}(\theta) = \frac{1}{D}\sum_i \frac{1}{2}\hat{\sigma}_i^{-2}||\mathbf{y}_i - \hat{\mathbf{y}}_i||^2 + \frac{1}{2}\log\hat{\sigma}_i^2$$

$$\sum_{k=1}^{K}\left(\mu_k - \mathrm{E}\left[Y_*^{(w,h)}|x_*;\hat{\Phi}\right]\right)^2 \mathrm{P}\left(Y_*^{(w,h)} \in B_k|\hat{\Phi}(x_*)\right)$$

CamVid	IoU
SegNet [28]	46.4
FCN-8 [29]	57.0
DeepLab-LFOV [24]	61.6
Bayesian SegNet [22]	63.1
Dilation8 [30]	65.3
Dilation8 + FSO [31]	66.1
DenseNet [20]	66.9
This work:	
DenseNet (Our Implementation)	67.1
+ Aleatoric Uncertainty	67.4
+ Epistemic Uncertainty	67.2
+ Aleatoric & Epistemic	**67.5**

(a) CamVid dataset for road scene segmentation.

NYUv2 40-class	Accuracy	IoU
SegNet [28]	66.1	23.6
FCN-8 [29]	61.8	31.6
Bayesian SegNet [22]	68.0	32.4
Eigen and Fergus [32]	65.6	34.1
This work:		
DeepLabLargeFOV	70.1	36.5
+ Aleatoric Uncertainty	70.4	37.1
+ Epistemic Uncertainty	70.2	36.7
+ Aleatoric & Epistemic	**70.6**	**37.3**

(b) NYUv2 40-class dataset for indoor scenes.

Table 1: **Semantic segmentation performance.** Modeling both aleatoric and epistemic uncertainty gives a notable improvement in segmentation accuracy over state of the art baselines.

Make3D	rel	rms	$\log_{10}$
Karsch et al. [33]	0.355	9.20	0.127
Liu et al. [34]	0.335	9.49	0.137
Li et al. [35]	0.278	7.19	0.092
Laina et al. [26]	0.176	4.46	0.072
This work:			
DenseNet Baseline	0.167	3.92	0.064
+ Aleatoric Uncertainty	**0.149**	3.93	**0.061**
+ Epistemic Uncertainty	0.162	**3.87**	0.064
+ Aleatoric & Epistemic	**0.149**	4.08	0.063

(a) Make3D depth dataset [25].

NYU v2 Depth	rel	rms	$\log_{10}$	δ_1	δ_2	δ_3
Karsch et al. [33]	0.374	1.12	0.134	-	-	-
Ladicky et al. [36]	-	-	-	54.2%	82.9%	91.4%
Liu et al. [34]	0.335	1.06	0.127	-	-	-
Li et al. [35]	0.232	0.821	0.094	62.1%	88.6%	96.8%
Eigen et al. [27]	0.215	0.907	-	61.1%	88.7%	97.1%
Eigen and Fergus [32]	0.158	0.641	-	76.9%	95.0%	98.8%
Laina et al. [26]	0.127	0.573	0.055	81.1%	95.3%	98.8%
This work:						
DenseNet Baseline	0.117	0.517	0.051	80.2%	95.1%	98.8%
+ Aleatoric Uncertainty	0.112	0.508	0.046	81.6%	95.8%	98.8%
+ Epistemic Uncertainty	0.114	0.512	0.049	81.1%	95.4%	98.8%
+ Aleatoric & Epistemic	**0.110**	**0.506**	**0.045**	**81.7%**	**95.9%**	**98.9%**

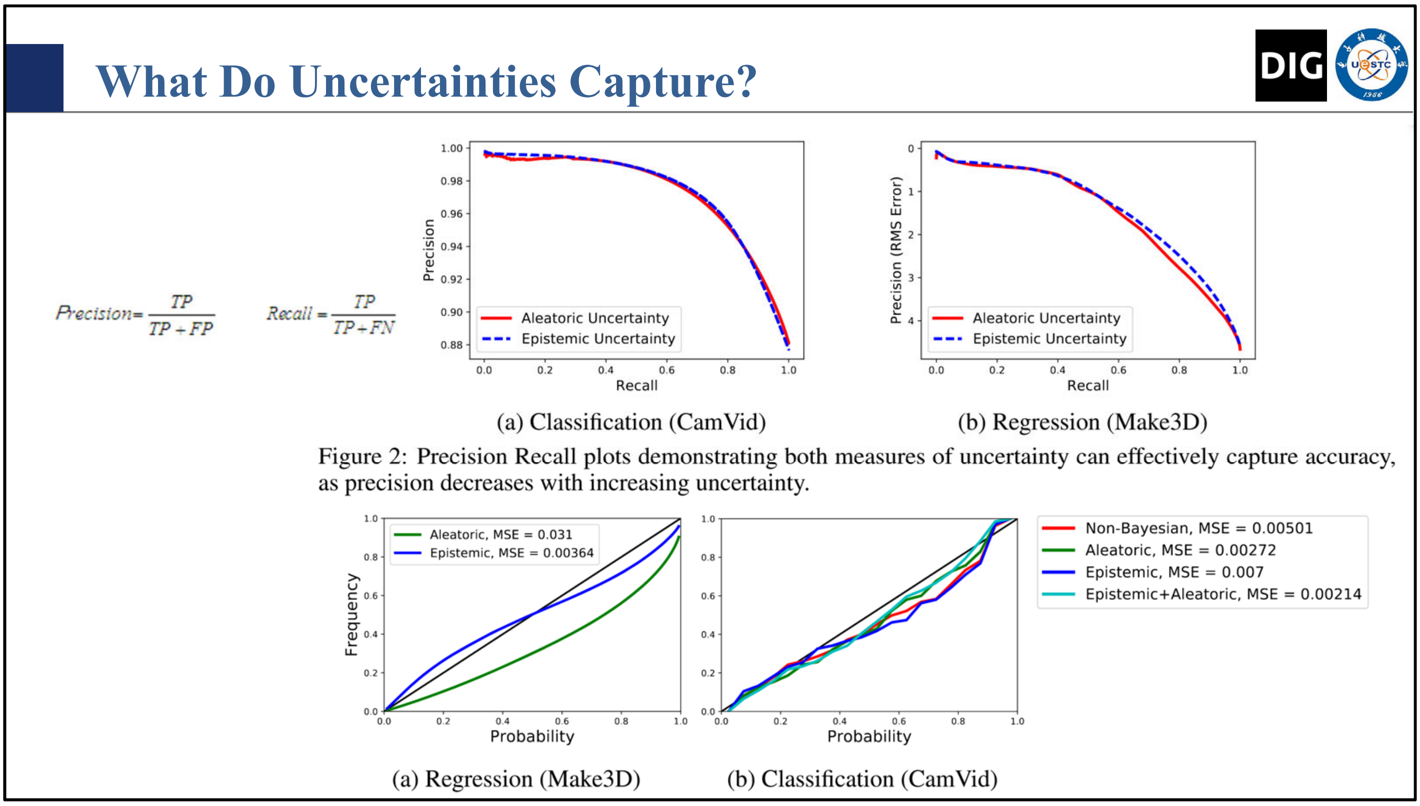
(b) NYUv2 depth dataset [23].

Table 2: **Monocular depth regression performance.** Comparison to previous approaches on depth regression dataset NYUv2 Depth. Modeling the combination of uncertainties improves accuracy.

ConOR +

Method	Prediction		
	rmse↓	rel↓	δ_1 ↑
MCC [6, 25]	3.011	0.081	0.915
BC [86]	2.878	0.078	0.919
GL+MCD [28]	3.337	0.102	0.875
GL+DE [51]	2.900	0.089	0.908
GL+WBS	3.064	0.083	0.906
GL+MBS	3.064	0.083	0.906
LGL+MCD [28]	3.219	0.158	0.836
LGL+DE [51]	2.852	0.132	0.873
LGL+WBS	2.965	0.132	0.870
LGL+MBS	2.965	0.132	0.870
ConOR+WBS	**2.709**	**0.075**	**0.928**
ConOR+MBS	**2.709**	**0.075**	**0.928**

例如，在图 5 和 6 的定性结果中，我们观察到图像中的大深度、反射面和遮挡边界的任意不确定性更大。



What Do Uncertainties Capture?

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN}$$

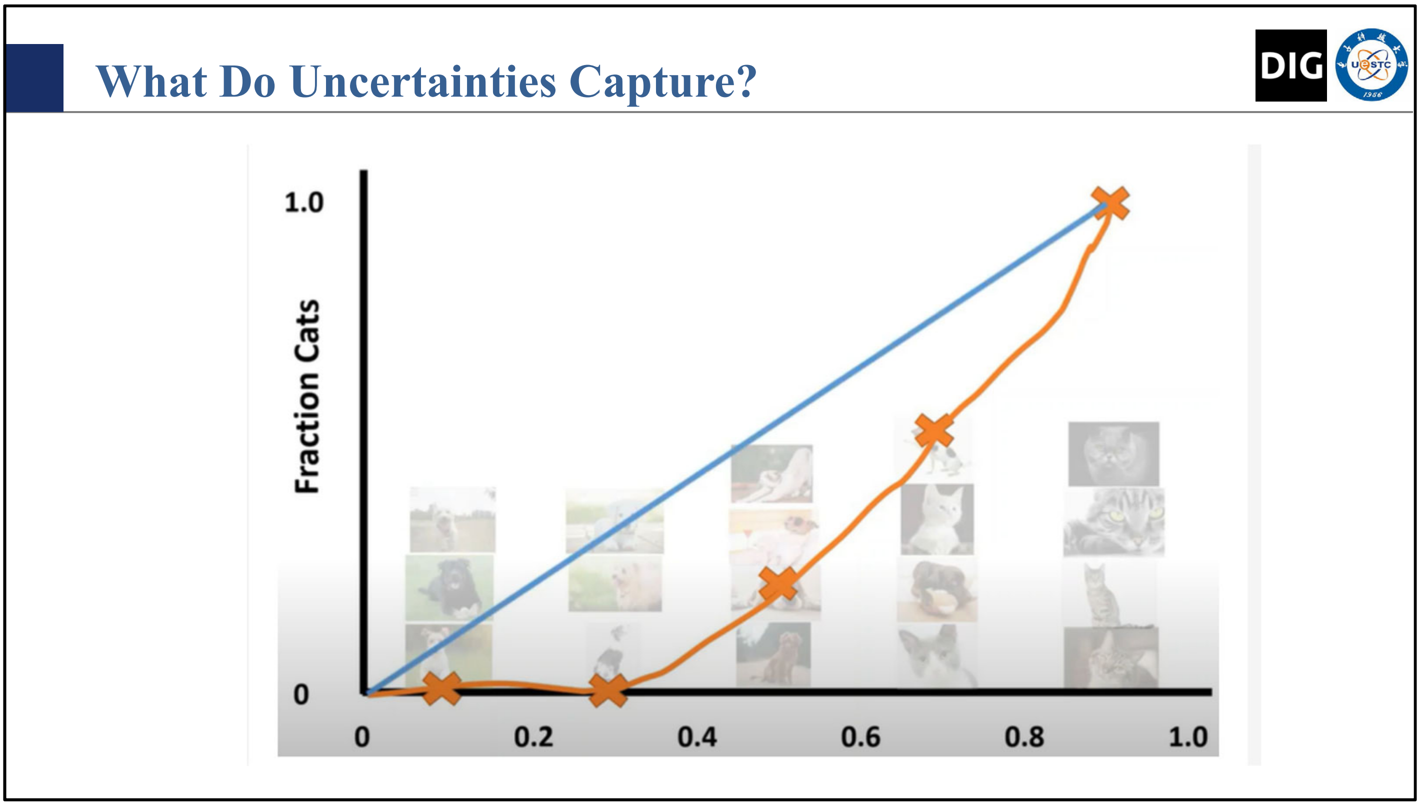
(a) Classification (CamVid) (b) Regression (Make3D)

Figure 2: Precision Recall plots demonstrating both measures of uncertainty can effectively capture accuracy, as precision decreases with increasing uncertainty.

(a) Regression (Make3D) (b) Classification (CamVid)

宁可把垃圾邮件标记为正常邮件(FN)，提高Precsion
每个不确定性对像素置信度的排名与其他不确定性类似。这表明，当仅对一种不确定性进行显式建模时，它会在可能的情况下尝试补偿替代不确定性的缺乏。

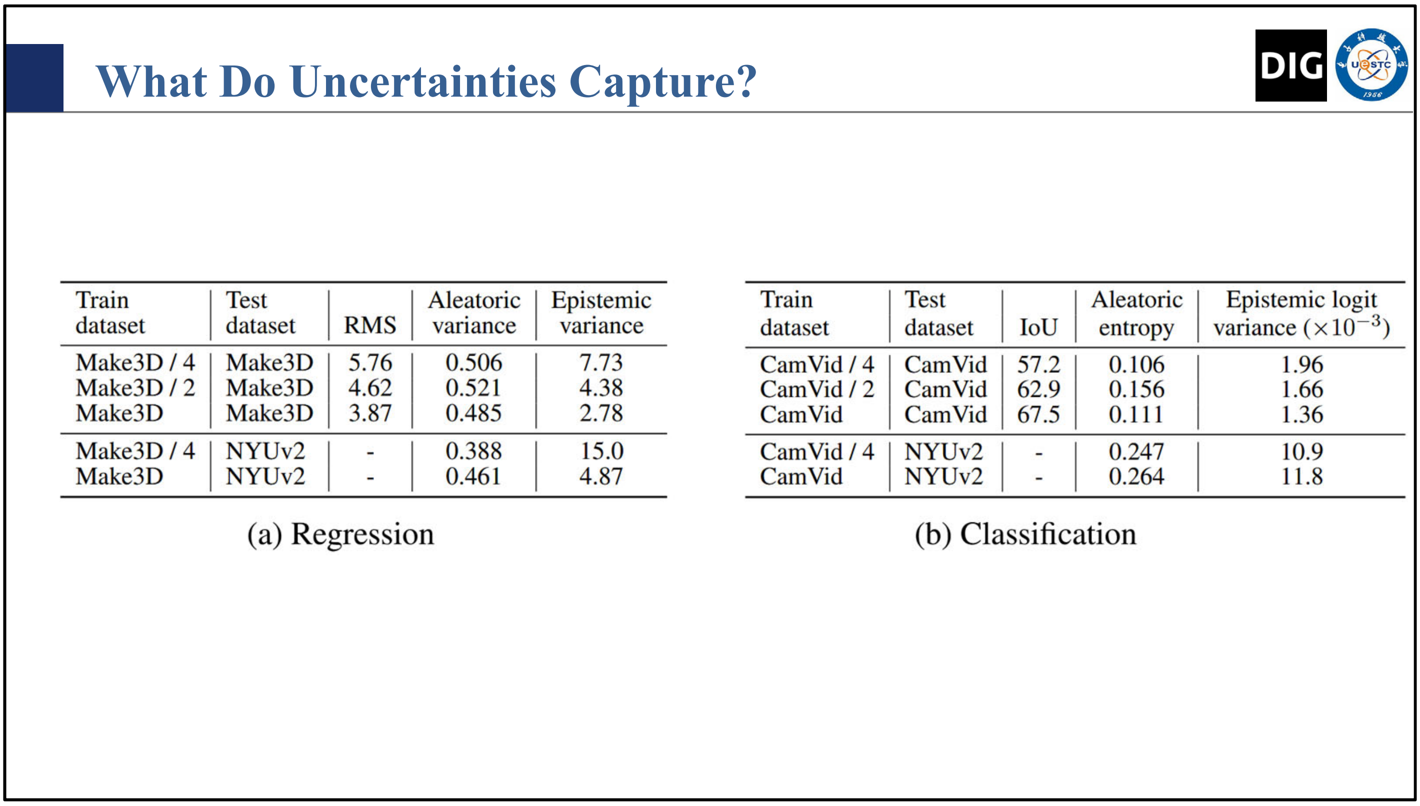
具体来说，分类器的预测是被认为是经过校准的，如果对于所有可能的概率阈值，给定该阈值的预测概率，模型预测正例的比例与实际观察到的正例比例相同。例如，如果我们从预测为正例概率为80%的所有实例中随机选择一个实例，那么我们期望该实例实际上为正例的概率也是80%。



What Do Uncertainties Capture?

宁可把垃圾邮件标记为正常邮件(FN)，提高Precsion
每个不确定性对像素置信度的排名与其他不确定性类似。这表明，当仅对一种不确定性进行显式建模时，它会在可能的情况下尝试补偿替代不确定性的缺乏。

具体来说，分类器的预测是被认为是经过校准的，如果对于所有可能的概率阈值，给定该阈值的预测概率，模型预测正例的比例与实际观察到的正例比例相同。例如，如果我们从预测为正例概率为80%的所有实例中随机选择一个实例，那么我们期望该实例实际上为正例的概率也是80%。



What Do Uncertainties Capture?

Train dataset	Test dataset	RMS	Aleatoric variance	Epistemic variance
Make3D / 4	Make3D	5.76	0.506	7.73
Make3D / 2	Make3D	4.62	0.521	4.38
Make3D	Make3D	3.87	0.485	2.78
Make3D / 4	NYUv2	-	0.388	15.0
Make3D	NYUv2	-	0.461	4.87

(a) Regression

Train dataset	Test dataset	IoU	Aleatoric entropy	Epistemic logit variance ($\times 10^{-3}$)
CamVid / 4	CamVid	57.2	0.106	1.96
CamVid / 2	CamVid	62.9	0.156	1.66
CamVid	CamVid	67.5	0.111	1.36
CamVid / 4	NYUv2	-	0.247	10.9
CamVid	NYUv2	-	0.264	11.8

(b) Classification

这表明ConOR（任意不确定性）估计的误差方差已经可以解释大部分预测不确定性，并且我们的方法可以使用重采样方法进一步增强对不确定性的理解。这个结果是合理的，因为 KITTI [30] 和 NYUv2 [77] 训练集的样本量较大导致较低。

ConOR > OR基于有效条件分布进行统计推断的显着性

Thanks