UNO: Uncertainty-aware Noisy-Or Multimodal Fusion for Unanticipated Input Degradation

Junjiao Tian, Wesley Cheung, Nathan Glaser, Yen-Cheng Liu and Zsolt Kira
Georgia Institute of Technology
{jtian73,wcheung8,nglaser3,ycliu,zkira}@gatech.edu

***Abstract*— The fusion of multiple sensor modalities, especially through deep learning architectures, has been an active area of study. However, an under-explored aspect of such work is whether the methods can be robust to degradations across their input modalities, especially when they must generalize to degradations not seen during training. In this work, we propose an *uncertainty-aware* fusion scheme to effectively fuse inputs that might suffer from a range of known and unknown degradations. Specifically, we analyze a number of uncertainty measures, each of which captures a different aspect of uncertainty, and we propose a novel way to fuse degraded inputs by scaling modality-specific output softmax probabilities. We additionally propose a novel data-dependent spatial temperature scaling method to complement these existing uncertainty measures. Finally, we integrate the uncertainty-scaled output from each modality using a probabilistic noisy-or fusion method. In a photo-realistic simulation environment (AirSim), we show that our method achieves significantly better results on a semantic segmentation task, compared to state-of-art fusion architectures, on a range of degradations (e.g. fog, snow, frost, and various other types of noise), some of which are unknown during training. We specifically improve upon the state-of-art[30] by 28% in mean IoU on various degradtions.**

I. INTRODUCTION

Image-based scene understanding methods for robotics, such as object detection and semantic segmentation, have been extensively studied and steadily improved in the past few years. The use of multiple sensing modalities on robots is common, however, and therefore there is an increasing interest in leveraging additional sensor information to complement image data. For example, depth information can help better separate objects that are hard to distinguish based on textures and color.

Utilizing multiple modalities entails fusion of different sensor streams that potentially provide complementary information. For example, depth estimation typically degrades quickly with distance, either in accuracy or resolution. Many works have explored where to fuse modality-specific streams topologically [27, 9, 30]. In general, researchers have attempted different fusion schemes such as early, late and hierarchical fusion schemes. Many works have also explored fusion schemes at different levels of representation in order to increase the interaction of different modalities [9, 30].

However, the variety of scenes and degradation in the real world presents a challenge to all fusion schemes. The ability to automatically adapt to a changing environment is the key to safety in application such as robotics and autonomous driving. A robust fusion scheme should dynamically adapt to sensor failure and noise, emphasizing the modalities that are less corrupted and more informative. Various works with different gating and attention mechanisms [21, 31, 30] have demonstrated the importance of weighting different modalities depending on the scene. Yet, there has been few works on adapting to a variety of degradation and noise. This is especially true for addressing degradations that do not appear in the training data, i.e. that are unknown *a-priori*.

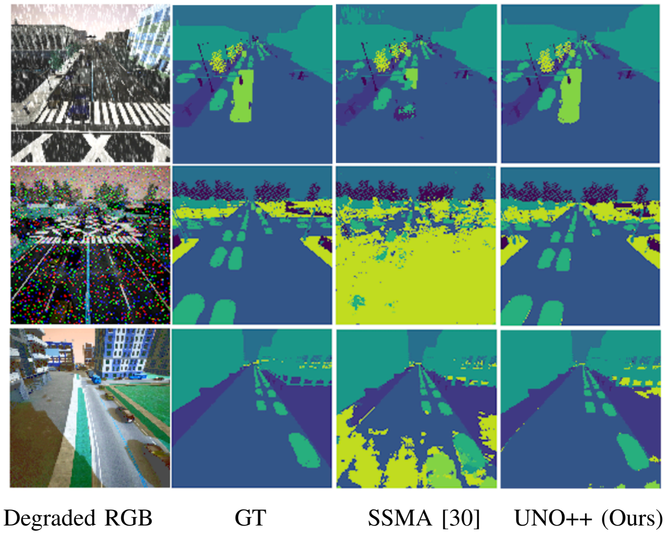

Fig. 1: **Performance of state-of-the-art fusion model and ours on degraded RGB data**. First row: Snow. Second row: Impulse noise. Third row: Brightness degradation. Note that the depth channel is not degraded.

In this paper, we investigate an adaptive fusion scheme for unseen degradations with application to RGB-D semantic segmentation. We leverage recent development in uncertainty estimation for deep neural networks [6, 15] and show that different uncertainty measures correlate differently to different types of degradations. We therefore propose a method to combine multiple types of uncertainties by representing their deviation from the training set (*deviation ratio)*, and use this criteria in a novel way to *calibrate* the output prediction probabilities. We further add an additional novel data-dependent spatial temperature scaling that models a spatial type of uncertainty not covered by existing approaches. Given uncertainty-modulated outputs from each modality, we finally propose a simple but flexible uncertainty-aware probabilistic fusion method with no learning parameters and show robust performance across different degradation. We show using a photorealistic simulator (AirSim [28]) across a variety of conditions such as fog, snow, and various types of noise, that our method achieves stronger fusion results than current state-of-art. Our improvement is especially noticeable in the cases where the specific degradations are not represented in the training set. Fig. 1 shows examples of performance of the state-of-the-art model SSMA [30] and our proposed method. Our method achieves a relative improvement in mean IoU of 11% over our strong baselines and 28% over state-of-art fusion methods.

The contributions of this paper are as follows:

- We propose a method for combining several uncertainty metrics, which capture different aspects of uncertainty, using a *deviation ratio* that encodes how the metrics deviate from the training set.
- We introduce an additional uncertainty method, the spatial temperature network, which captures a data-dependent spatial uncertainty that is absent in the existing uncertainty metrics.
- We propose a probabilistic uncertainty-aware fusion scheme, Uncertainty-aware Noisy-Or (UNO), that dynamically adapts to the changing environment by combining an arbitrary set of experts (e.g. modalities or architectures). Our method has several advantages, including speed (i.e., no training needed) and the ability to dynamically fuse an arbitrary (potentially changing) set of modalities.
- We demonstrate significantly increased robustness of our method, compared to state-of-art fusion baselines, on a photo-realistic simulation-based dataset across a range of degradations, including ones that are not present in the training set.

II. RELATED WORK

Fusion Architectures for RGB-D Semantic Segmentation A number of fusion architectures have been developed for combining modalities [27, 9, 30], and recently it has been shown that variations of attention and gating mechanisms [16, 2, 32] can adapt to dynamic environments by weighting modalities differently for conditions that occur in the training data. While most works [21, 31] have considered external environmental degradations such as rain, snow, glare, low-lighting, and seasonal appearance changes, more recent works [17] address robustness to different types of internal image degradations such as Gaussian noise. However, we observe that these methods are both trained and tested on the same subset of degradations–the test set noises are in the training distribution and hence explicitly learn-able by the network. Although [12] include ways to augment data by applying many of the common corruptions and perturbations, we posit that it is unrealistic to anticipate every degradation that may be encountered. Hence, our methods attempt to address this key limitation of current work, namely their inability to reliably and feasibly address all possible forms of degradations. For more details on the recent trends and architectures for multimodal fusion, we refer to [26].

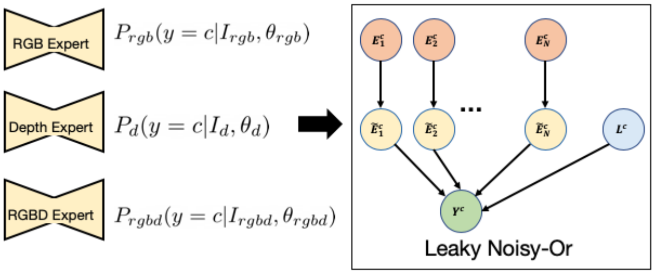

Fig. 2: **Fusing experts with Noisy-Or**. Each modality-specific expert generates an uncertainty-based distribution. Final prediction is obtained through a (leaky) Noisy-Or fusion.

Noisy-Or Approximate Bayesian Inference We model the fusion process as Bayesian inference. Under certain cause-independent assumptions [11], the conditional probability in a Bayesian network can be approximated by a Noisy-Or gate [25]. Unlike logical-or, Noisy-Or is more realistic because each parent has a non-negligible probability of being inhibited. Practically for fusion, this means that no hard threshold is needed and predictions from each modality are considered with non-negligible probability. This framework has been expanded to include a leak probability which accounts for causes not covered by all of the independent parents [13].

Uncertainty Estimation We are interested in an uncertainty estimator that correlates well with the degree of anticipated and unanticipated degradation. Regarding unanticipated degradation, there are many works on detecting misclassification and out-of-distribution (OOD) data by measuring certain notions of uncertainty or confidence. It has been noted that out-of-distribution images can be identified through epistemic uncertainties [15]. Many other per-pixel uncertainty measures have also been developed and compared in [1] including ODIN [19], Bayesian networks [14, 23], density estimation [4], and OOD training [3].

In particular, Bayesian networks [14, 23] have been shown to yield desirable properties for modeling uncertainty, such as a model confidence that correlates with accuracy. Their methods involve using Monte-Carlo dropout (MCDO) [7] as the primary means of approximate inference, with several measures such as predictive entropy and mutual information that can be calculated from these sampling passes.

We show in our scenarios that model uncertainty (approximated by dropout) does not correlate with all degradations, and calibration methods such as temperature scaling [8], must be trained on the specific degradations that will be encountered. In our work we propose a novel *uncertainty-based* calibration as well as a data-dependent spatial form of temperature scaling; we then combine multiple notions of uncertainty to maximize robustness.

III. METHOD

In this section, we introduce the Uncertainty-aware Noisy-Or (UNO) fusion scheme. We briefly highlight three conventional uncertainty metrics, each of which captures a different element of uncertainty, and our proposed *deviation ratio* in Section III-A that allows us to combine these

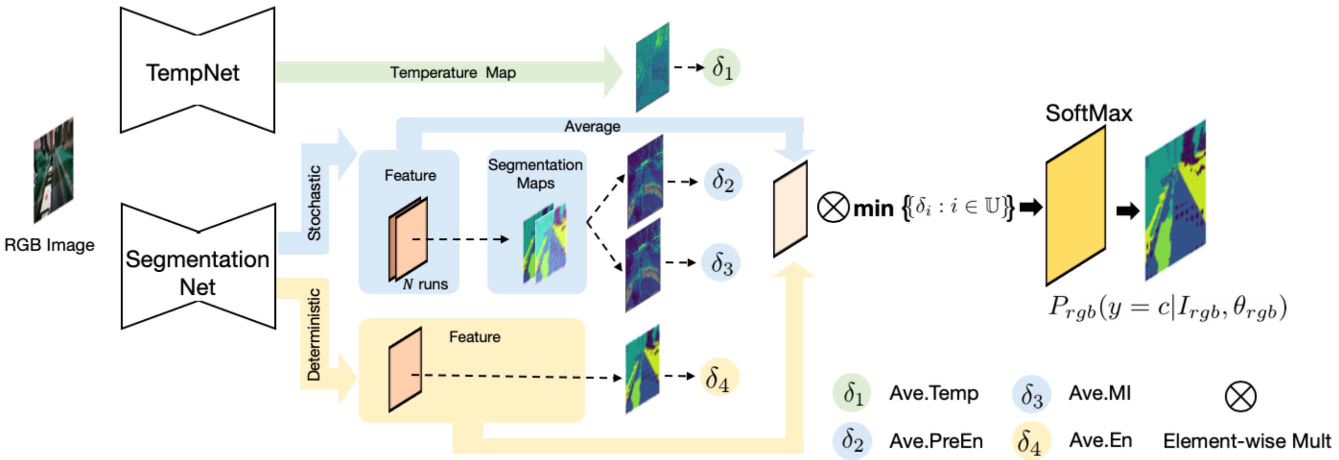

Fig. 3: **Overall pipeline for a single expert (RGB branch)**. We combine multiple uncertainty metrics using the deviation ratios (δs) calculated for a test sample using different uncertainty metrics. Deviation ratio is defined in Section III-A. Either stochastic model or deterministic model is used at a time. The performance of both is examined in Section IV-C.

uncertainty metrics. We then describe a novel learning-based uncertainty metric, *TempNet* (see Section III-B) for capturing a data-conditioned spatial uncertainty that is not covered by the existing approaches. Finally, we propose a probabilistic framework for mutli-modal fusion in Section III-C.

In the discussions below, we define $\mathbb{C}$ as the set of classes we are interested in classifying, $\mathbb{U}$ as the set of employed uncertainty metrics. We view each modality-specific segmentation networks as different experts and denote $\{E_1, .., E_i\} \in \mathbb{E}$ as the set of independent segmentation networks (experts).

A. Uncertainty Estimation and Deviation Ratio

A number of methods exist for producing uncertainty estimates over the output predictions of a neural network [8, 6, 15], namely predictive entropy, mutual information, and deterministic entropy. While many works focus on the accuracy of such estimates, we propose to use them to re-weight modalities during fusion. Unlike hand-chosen weighting methods [24] that use uncertainty, we propose to modulate modalities in a novel way through automatic *scaling* of the outputs scores (and hence softmax probabilities); this procedure is similar to calibration methods [8] but conditioned on uncertainty. The goal of uncertainty scaling is to "soften" the softmax probabilities depending on how different (i.e. out-of-distribution) the uncertainties are from the training data.

We first describe the existing uncertainty metrics used in this paper, as proposed by [6]. The **predictive entropy** of a predictive distribution, $\mathbb{H}[y|x, D_{train}]$, given a test sample x and training data D_{train} can be approximated by collecting outputs from T stochastic forward passes with different dropout samples through the network (i.e. MCDO):

$$\hat{\mathbb{H}}[y|x, D_{train}] \approx -\sum_{c}^{\mathbb{C}} \left(\frac{1}{T} \sum_{t} p(y = c|x, \hat{\theta}_t) \log \frac{1}{T} p(y = c|x, \hat{\theta}_t) \right), \quad (1)$$

where c is over all classes and $p(y = c|x, \hat{\theta}_t)$ is the probability mf class c given input x and $\hat{\theta}_t$ is the sampled weights at stochastic pass t.

The **mutual information** can be approximated with a similar procedure:

$$\hat{\mathbb{I}}[y, w|x, D_{train}] \approx \hat{\mathbb{H}}[y|x, D_{train}] + \frac{1}{T} \sum_{c,t} p(y = c|x, \hat{\theta}_t) \log p(y = c|x, \hat{\theta}_t), \quad (2)$$

We also compare with the **entropy** of a deterministic model:

$$\mathbb{H}[y|x, D_{train}] = -\sum_{c} p(y = c|x, \theta) \log p(y = c|x, \theta), \quad (3)$$

where w is the model's learned parameters and is fixed at inference (note only one forward pass is required).

To automatically scale the output probabilities conditioned on the degradation level, a new metric that scales dynamically with uncertainty is needed. We propose to capture how deviated a test sample is from the training distribution using uncertainty metrics. We define a deviation ratio, which reports a numeric score less than unity if a test sample is out-of-distribution and unity if the sample is in-distribution (note that intuitively degradations should increase uncertainty). Specifically:

$$\delta = \frac{\mu_{train}}{max\left(0, \mu_{test} - \mu_{train} - \sigma_{train}\right) + \mu_{train}}, \quad (4)$$

where μ_{train} is the training average of a specific uncertainty metric aggregated across all images and averaged over pixels in the training set and σ_{train} is the standard deviation of the uncertainty metric scores. μ_{test} is the average uncertainty score for a test sample. The uncertainty metrics provide pixel-wise scores, which we average over an entire image. We perform this averaging step because the per-pixel uncertainty metrics can be unreliable as a local indicator of deviation, as shown by [1]. Thus, from the three uncertainty metrics listed above, three deviation ratios can be calculated:

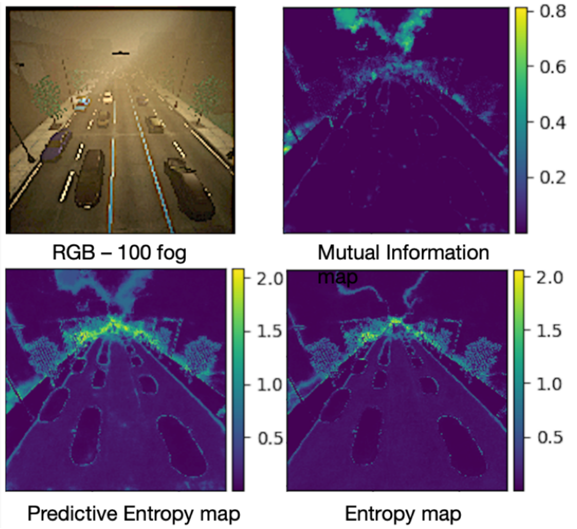

Fig. 4: **RGB uncertainty measurements with significant fog degradation**. Conventional uncertainty metrics do not capture the spatial degradation caused by fog.

average mutual information (**Ave.MI**), average predictive entropy (**Ave.PreEn**) and average entropy (**Ave.En**).

In the end, we combine the deviation ratios of different metrics using a Min operation. Intuitively, we choose the metric that is most sensitive to the current degradation, adopting a conservative selection method in which we assume worst-case for a model's uncertainty.

$$\delta_{min} = \min\left([\delta_i : i \in \mathbb{U}]\right), \tag{5}$$

To calibrate a network to reflect uncertainty in the presence of degradation, we define the uncertainty-calibrated prediction as follows.

$$p_i = \text{Softmax}\left([l_i^1, ..., l_i^c] * \delta_{min}\right), \tag{6}$$

where $[l_1^i, ..., l_c^i]$ are the pre-softmax logits. Thus, if a sample is far from the training distribution, δ will be a scalar less than 1 and thus "softens" the distribution and makes the model less confident in its prediction. The complete segmentation and scaling pipeline for the RGB branch is shown in Fig. 3.

For each modality-specific expert, a different δ_{min} is calculated. For the multimodal expert, we do not extract new deviation ratios; rather a second Min operation is performed on all the δ_{min}'s from involved modality-specific experts.

B. Spatial Temperature Network (TempNet)

In the calibration literature, conventional temperature scaling [8] is not able to adapt to different test conditions because it utilizes a scalar trained to a specific calibration set. Another observation is that degradation is often regional. For example, fog affects vision further into the distance and a good temperature model should be able to flatten the distributions for points more distant from the viewpoint.

We therefore introduce a spatial and data-conditioned temperature network to capture uncertainty induced by degradation. We show that the output of this model can be interpreted as an uncertainty which is not captured by conventional uncertainty extraction methods. As an example, Fig. 4 shows an example of fog degradation and corresponding uncertainty maps. These uncertainties capture uncertainty along edges of objects but do not capture spatial uncertainty.

TempNet is a shallow version of SegNet and uses 2 convolution and pooling/upsampling blocks for the decoder and encoder. Unlike prior methods, it is conditioned on the data, i.e. the input to the temperature network is the same as for the segmentation network and the output is a single-channel *spatial temperature map*, $T \in d_o \times d_o$. The average temperature deviation ratio (**Ave.Temp**) uses the average of the test spatial temperature map and applying Eq. (4).

To train TempNet, we minimize the Negative Log Likelihood of the correct class label for each pixel.

$$p_{ij} = \text{Softmax}\left([l_1, ..., l_c]_{ij} * t_{ij}\right), \tag{7}$$

$$L = -\sum_{ij} \log\left(p_{ij}(y = c|x, \theta)\right), \tag{8}$$

where t_{ij} is the i_{th} row and j_{th} column element of the temperature map T and l_{ij} is the pre-softmax logit of the segmentation output. The segmentation network is pretrained and kept fixed when training the temperature network. The same training procedure is done independently per modality.

C. Noisy-Or Fusion

We now introduce our proposed probabilistic fusion scheme that combines the uncertainty-modulated outputs of each modality. The final predictive distribution for a pixel is obtained by a *Noisy-Or operation* for each class and then normalized across all classes:

$$I(y = c) = 1 - \prod_i 1 - p_i(y = c|x_i, \theta_i) \quad \forall i \in \mathbb{E}, \tag{9}$$

$$p(y = c) = \frac{I_c}{\sum_j I_j} \quad \forall j \in \mathbb{C}, \tag{10}$$

where $p_i(y = c|x_i, \theta_i)$ is the predictive probability of expert i for class c, x_i and θ_i are the input and parameters of expert i and p_c is the final probability for class c.

Figure 2 illustrates the fusion process. In Bayesian networks, Noisy-Or can be used to model causality between N causes $E_1, E_2, ..., E_N$ and their common effect Y under certain independent causality assumptions: 1) Each of the causes is sufficient to produce the effect in the absence of all other causes, and 2) The ability of being sufficient is not affected by the presence of other causes. Practically Noisy-Or has some desirable properties for fusion. It preserves disagreement and accentuates agreement between experts. Another unique property of Noisy-Or compared to other methods such as multiplication, which simply multiplies two probabilities is, that when multiple experts give different predictive distributions to some discrete classes, it is possible that Noisy-Or selects the class on which one expert is really confident regardless of agreement on other classes while still being robust to outliers. A toy example demonstrates the flexibility of Noisy-Or in Fig. 5.

We argue that these independent causality assumptions are satisfied because each E_i is a complete segmentation

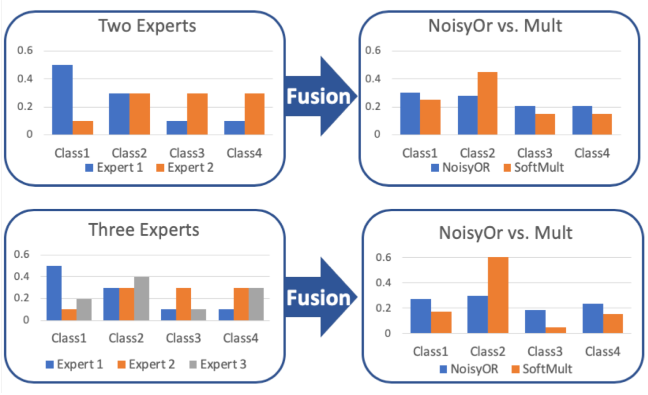

Fig. 5: **Noisy-Or vs. Multiplication 4 class prediction Example**. **Row 1**: Expert 1 is confident in class 1 while expert 2 is uncertain. Noisy-Or selects case 1 whereas Mult selects case 2 on which the two experts agree the most. **Row 2**: A third expert gives a confident prediction for class 2, and both Noisy-Or and Mult select case 2.

network capable of producing a probability $p_i(y = c|x_i, \theta_i)$ independent of other experts and is not affected by the presence of others. The leak node models the causes not covered by the independent experts. In our case, we adopt an off-the-shelf RGB-D fusion model as as the leak node.

The causality independence assumption makes this framework flexible. It is easy to add or remove new modality-specific experts and multimodal expert can be introduced as a leak node. By incorporating multiple uncertainty-aware modality-specific experts and multimodal experts into one framework, our Noisy-Or fusion model is robust to OOD.

IV. EXPERIMENTS

A. Dataset

We utilize the AirSim [28] simulator to collect a RGB-D semantic segmentation dataset. To collect this dataset, we fly a drone between road intersections in the AirSim City Environment, and we record the same trajectory under different weather settings, such as snow and fog, that can be varied on a scale from 0 (lowest intensity) to 100 (highest intensity). Note that a 100 setting does not correspond to complete snow whiteout or fog blackout. All models are trained on 0 and 50 fog levels. In order to evaluate both in- and out-of-distribution degradations, the models are tested on these same fog levels, in addition to a fog level of 100, frost, snow, and various degradation augmentation techniques investigated in [12]. For these augmentations, we use a severity of 3 (on a scale from 0 to 5) which equates to about a 50% decrease in average precision according to [22]. We randomly select the between-intersection segments to be a part of either training, testing, or validation sets. RGB and depth frames were captured at 512x512 pixels, and following the work of [5] we use a jet mapping to frame our depth inputs. Overall, our dataset contains 6857 labeled training images, 472 validation images and 359 testing images to which we apply both in- and out-of-simulation degradation.

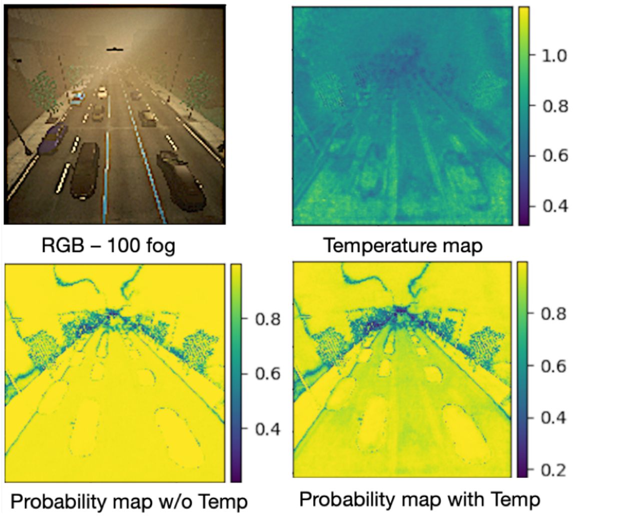

Fig. 6: **Effects of the temperature map on RGB softmax outputs**. Temperature map captures spatial uncertainty caused by fog (it is less confident in the foggy area). Smaller temperature (darker coloring) indicates lower confidence

B. Benchmarking

For comparison, we consider two state-of-art learned multimodal fusion schemes: FuseNet [9] and SSMA [30]. To make the SSMA framework comparable to FuseNet and UNO, we replace the original ResNet-50 encoder [10] to SSMA with a VGG 16-layer encoder [29] To measure overall performance of our segmentation networks, we use the mean intersection-over-union (mIoU) metric [20].

All models are trained for a maximum of 500K iterations using Adam [18] with a learning rate of 10^{-5} and default settings, and the best on the validation set is chosen for evaluation. When training the temperature map and scaling parameters, we use a two-step procedure [8]. We first train our network on the segmentation task with our training set. Then we find the optimal parameters for temperature for 100K iterations while the segmentation network is fixed.

C. Results

Fusion Performance In this section, we report the results of uncertainty-based Noisy-Or fusion with different deviation ratios as an ablation study in Table I and compare our adaptive fusion architecture to other fusion methods in Table II. Table I shows that MCDO-based uncertainties as a global indicator of OOD ratio do not outperform single-pass entropy and the **min** operation can effectively choose the most sensitive uncertainty depending on the degradation. Therefore, we use **min**(Ave.Temp,Ave.En) as the deviation ratio for our model in the following experiments because MCDO-based uncertainties require multiple passes with higher computation costs and longer runtime at inference. Table II compares our methods (**UNO** without SSMA as a leak expert and **UNO++** with) to other fusion methods. The results demonstrate that our method outperforms the state-of-the-art fusion models on unseen degradations and can be easily applied to any off-the-shelf multimodal fusion model to improve its performance

	In Distribution		RGB Degradation							Depth Degradation			
	0 fog	50 fog	100 fog	MotionBlur	Frost	Snow	Brightness	Blackout	Impulse	Gaussian	ShotNoise	Impulse	Average
1, Ave.Temp	80.75	77.79	75.73	73.36	75.59	74.89	72.47	78.71	78.74	68.22	72.34	64.86	74.45
2, Ave.En	82.62	79.70	77.34	81.21	80.37	79.50	78.32	79.65	79.43	77.55	79.32	47.62	76.89
3, Ave.PreEn	80.80	77.79	77.17	79.68	78.78	78.06	77.51	78.65	78.52	75.47	77.62	45.43	75.46
4, Ave.MI	80.81	77.80	78.07	78.28	78.14	78.05	78.30	73.41	76.69	75.76	77.62	47.84	75.06
min(1,2)	82.62	79.69	77.67	81.24	80.38	79.57	78.33	79.72	79.43	77.55	79.32	64.80	**78.36**
min(1,3,4)	80.77	77.75	78.14	78.29	78.35	77.94	78.09	78.53	78.52	75.79	77.64	63.53	76.95

TABLE I: **Ablation:** performance of Noisy-Or fusion with different deviation ratios (uncertainty metrics) on Mean IoU. Ave.Temp and Ave.En require a single deterministic pass whereas Ave.PreEn and Ave.MI require multiple MCDO passes.

	In Distribution		RGB Degradation							Depth Degradation			
	0 fog	50 fog	100 fog	MotionBlur	Frost	Snow	Brightness	Blackout	Impulse	Gaussian	ShotNoise	Impulse	Average
SoftMult	84.25	81.22	77.57	72.22	75.79	74.73	71.28	61.80	64.69	50.48	64.46	40.65	68.26
SoftMult(T)	84.24	81.24	76.88	71.90	74.90	74.88	70.37	59.93	62.86	52.36	65.78	43.46	68.23
NoisyOr	82.56	79.69	76.47	72.81	75.22	71.84	71.84	70.81	77.76	63.74	68.14	47.47	71.76
NoisyOr(T)	82.56	79.74	75.87	72.65	75.69	73.85	71.13	69.13	77.19	64.04	68.52	49.19	71.63
FuseNet [9]	85.41	80.61	80.98	73.24	71.59	69.33	70.85	49.94	54.55	3.37	5.58	4.40	50.89
SSMA [30]	87.35	82.89	82.58	73.94	69.88	65.80	70.90	33.02	34.15	51.08	55.35	42.93	62.49
UNO	82.62	79.69	77.67	81.24	80.38	79.57	78.33	79.72	79.43	77.55	79.32	64.80	78.36
UNO++	86.70	83.51	83.23	83.11	82.33	81.70	80.37	79.13	79.79	78.28	79.86	61.97	**80.00**

TABLE II: **Comparison:** performance of UNO and UNO++ against other non-learning and learning baselines on Mean IoU. SoftMult(T) and NoisyOr(T) use the original temperature scaling [8].

	In Distribution		RGB Degradation					Depth Degradation		
RGB/D	0 fog	50 fog	100 fog	MotionBlur	Brightness	Blackout	Impulse	Gaussian	ShotNoise	Impulse
1, Ave.Temp	1.00/1.00	1.00/1.00	0.97/1.00	1.00/1.00	0.82/1.00	0.03/1.00	0.05/1.00	1.00/0.88	1.00/0.93	1.00/0.75
2, Ave.En	1.00/1.00	1.00/1.00	0.97/1.00	0.54/1.00	0.48/1.00	0.46/1.00	0.26/1.00	1.00/0.47	1.00/0.32	1.00/1.00
3, Ave.PreEn	1.00/1.00	1.00/1.00	0.79/1.00	0.39/1.00	0.41/1.00	0.51/1.00	0.35/1.00	1.00/0.50	1.00/0.30	1.00/1.00
4, Ave.MI	1.00/1.00	1.00/1.00	0.42/1.00	0.13/1.00	0.18/1.00	0.88/1.00	1.00/1.00	1.00/0.42	1.00/0.22	1.00/1.00

TABLE III: **Sensitivity:** average test deviation ratio using different uncertainty metrics for in/out of distribution conditions.

across degradations. Also, our baseline Noisy-Or shows better results than SSMA under unseen degradations. This shows that multimodal expert is less robust when any of its input modality is degraded in a manner not known *a-priori*. Note also that normal temperature scaling does not provide additional improvement across degraded conditions.

Temperature Maps We qualitatively show that conventional uncertainty metrics such as predictive entropy and mutual information from multiple MCDO stochastic passes or entropy from a single deterministic pass fail at detecting uncertainty associated with degradation. The temperature map, on the other hand, captures the spatial degradation as shown in Fig. 6. As shown in table III, when the temperature map is used as a global scaling deviation ratio it is sensitive to various degradation and especially when there is *Impulse noise* or *blackout* degradation on the RGB or depth channel. We hypothesize that TempNet learns data-dependent spatial uncertainty due to spatial noise through training whereas other uncertainties extract pixel-wise statistical uncertainty based on predictive distributions.

Uncertainty and Degradation In this section, we report the average test deviation ratios calculated from all aforementioned uncertainty metrics including average temperature for in-distribution and different unseen degraded conditions. As shown in table III, we list the average deviation ratio for in-distribution conditions, i.e., 0 fog and 50 fog conditions, and various degradation on RGB input or depth input respectively. The table shows that all metrics report a deviation ratio of unity on average for in-distribution data and mostly less than unity for OOD inputs. However, they exhibit different sensitivity. For example, average temperature is sensitive to blackout and impulse noise and not as responsive for motion blur and brightness degradation on the RGB channel. On the contrary, MCDO uncertainties and entropy are inactive to *Impulse Noise*, justifying our combination of uncertainties.

V. CONCLUSION AND FUTURE WORK

We have presented an adaptive framework for multimodal fusion that, unlike existing fusion methods, addresses unanticipated, out-of-training degradations. We benchmark different measures of uncertainty and propose a novel uncertainty-based softmax scaling as well as a deviation ratio for combining uncertainty metrics. We also propose a new data-conditioned spatial uncertainty (*TempNet*) and a simple but effective noisy-or fusion scheme that can combine an arbitrary number of modalities. Results show superior performance to existing state-of-art and extensibility for incorporating them as additional experts. Next steps include experimentation with additional uncertainty metrics and analysis of their trade-offs.

References

[1] Hermann Blum et al. "The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation". In: *arXiv e-prints* (Apr. 2019).

[2] Liuyuan Deng et al. "RFBNet: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation". In: *arXiv preprint arXiv:1907.00135* (2019).

[3] Terrance DeVries and Graham W. Taylor. "Learning Confidence for Out-of-Distribution Detection in Neural Networks". In: *arXiv e-prints*, arXiv:1802.04865 (Feb. 2018), arXiv:1802.04865. arXiv: 1802.04865 [stat.ML].

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *arXiv e-prints*, arXiv:1605.08803 (May 2016), arXiv:1605.08803. arXiv: 1605.08803 [cs.LG].

[5] Andreas Eitel et al. "Multimodal Deep Learning for Robust RGB-D Object Recognition". In: *arXiv e-prints* (July 2015).

[6] Yarin Gal. "Uncertainty in deep learning". PhD thesis. PhD thesis, University of Cambridge, 2016.

[7] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *arXiv e-prints* (June 2015).

[8] Chuan Guo et al. "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1321–1330.

[9] C. Hazirbas et al. "FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture". In: *Asian Conference on Computer Vision*. Nov. 2016.

[10] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *arXiv e-prints* (Dec. 2015).

[11] David Heckerman and John S Breese. "A new look at causal independence". In: *Uncertainty Proceedings 1994*. Elsevier, 1994, pp. 286–292.

[12] Dan Hendrycks and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *arXiv e-prints*, arXiv:1903.12261 (Mar. 2019), arXiv:1903.12261. arXiv: 1903.12261 [cs.LG].

[13] Max Henrion. "Some Practical Issues in Constructing Belief Networks." In: *UAI*. Vol. 3. 1987, pp. 161–173.

[14] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding". In: *arXiv preprint arXiv:1511.02680* (2015).

[15] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems*. 2017, pp. 5574–5584.

[16] Jaekyum Kim et al. "Robust Deep Multi-modal Learning Based on Gated Information Fusion Network". In: *arXiv e-prints*, arXiv:1807.06233 (July 2018), arXiv:1807.06233. arXiv: 1807.06233 [cs.CV].

[17] Taewan Kim and Joydeep Ghosh. "On Single Source Robustness in Deep Fusion Models". In: *arXiv e-prints* (June 2019).

[18] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv e-prints*, arXiv:1412.6980 (Dec. 2014), arXiv:1412.6980. arXiv: 1412.6980 [cs.LG].

[19] Shiyu Liang, Yixuan Li, and R. Srikant. "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks". In: *arXiv e-prints* (June 2017).

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[21] Oier Mees, Andreas Eitel, and Wolfram Burgard. "Choosing smartly: Adaptive multimodal fusion for object detection in changing environments". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 151–156.

[22] Claudio Michaelis et al. "Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming". In: *arXiv e-prints* (July 2019).

[23] Jishnu Mukhoti and Yarin Gal. "Evaluating Bayesian Deep Learning Methods for Semantic Segmentation". In: *arXiv e-prints*, arXiv:1811.12709 (Nov. 2018), arXiv:1811.12709. arXiv: 1811.12709 [cs.CV].

[24] G. Papandreou et al. "Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition". In: *2007 IEEE 9th Workshop on Multimedia Signal Processing*. Oct. 2007, pp. 264–267. DOI: 10.1109/MMSP.2007.4412868.

[25] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[26] D. Ramachandram and G. W. Taylor. "Deep Multimodal Learning: A Survey on Recent Advances and Trends". In: *IEEE Signal Processing Magazine* 34.6 (Nov. 2017), pp. 96–108. ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2738401.

[27] Joel Schlosser, Christopher K Chow, and Zsolt Kira. "Fusing lidar and images for pedestrian detection using convolutional neural networks". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016, pp. 2198–2205.

[28] Shital Shah et al. "AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles". In: *Field and Service Robotics*. 2017. eprint: arXiv:1705.05065. URL: https://arxiv.org/abs/1705.05065.

[29] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[30] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. “Self-Supervised Model Adaptation for Multimodal Semantic Segmentation”. In: *arXiv e-prints*, arXiv:1808.03833 (Aug. 2018), arXiv:1808.03833. arXiv: `1808.03833 [cs.CV]`.

[31] Abhinav Valada et al. “AdapNet: Adaptive semantic segmentation in adverse environmental conditions”. In: May 2017, pp. 4644–4651. DOI: `10.1109/ICRA.2017.7989540`.

[32] Jin Zeng et al. “Deep Surface Normal Estimation with Hierarchical RGB-D Fusion”. In: *arXiv e-prints* (Apr. 2019).