Poster Session

NEURAL INFORMATION PROCESSING SYSTEMS

L-TTA: Lightweight Test-Time Adaptation Using a Versatile Stem Layer

Jin Shin, Hyun Kim*

Seoul National University of Science and Technology, Seoul, Korea
{shinjin0103, hyunkim}@seoultech.ac.kr

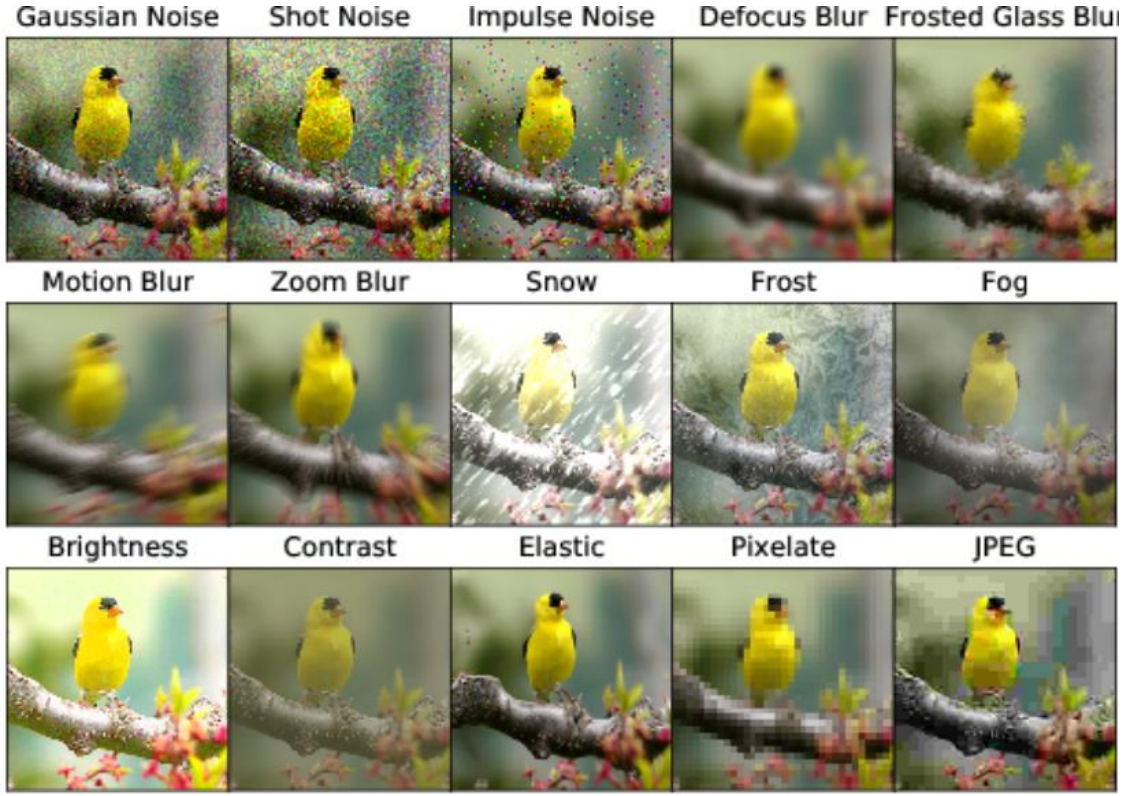
IDSL
Intelligent Digital Systems Design Lab.

SEOULTECH

Introduction - 1

- **Test-Time Adaptation (TTA):**

  - TTA adapts pre-trained models to unseen distribution of data during inference, using only input (*i.e.*, images) without labels by fine-tuning some layers.



Effect of different test samples in test time entropy minimization

(Efficient Test-Time Model Adaptation without Forgetting, PMLR, 2022)

Introduction - 2

- **Motivations:**

  Recent TTA methodologies have focused on minimizing entropy (EM), but this approach **requires forward/backward process** and is limited in terms of data leveraging.

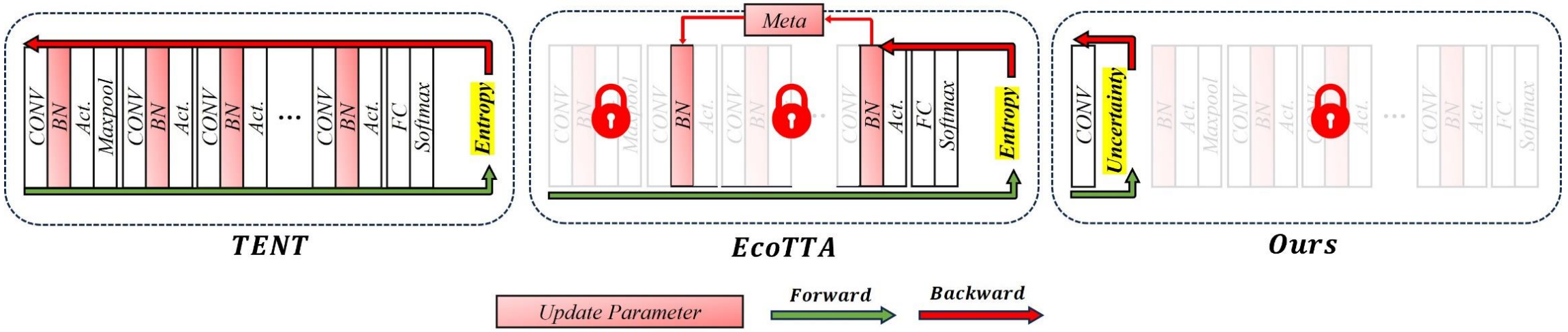
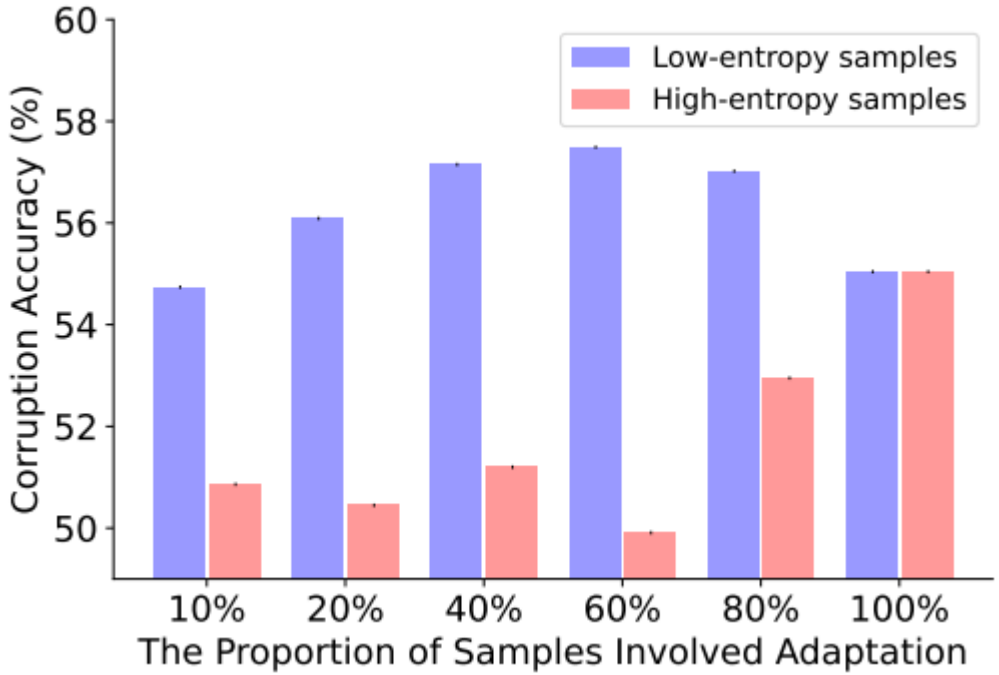


Diagram comparing the forward/backward flow and update process

Introduction - 3

- **Motivations:**

  Recent TTA methodologies have focused on minimizing entropy (EM), but this approach requires forward/backward process and is **limited in terms of data leveraging**.



Effect of different test samples in test time entropy minimization
(Efficient Test-Time Model Adaptation without Forgetting, PMLR, 2022)

Introduction - 4

- **Goals:**

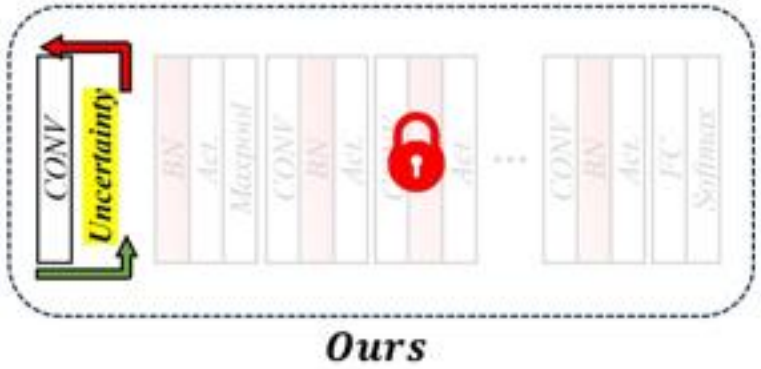
  1. Practicality in training: Minimizing the resources required for training such as memory, **to attain acceptable a reasonable prediction accuracy** in a D_t.

  2. Scalability: Designing to be non-invasively and conveniently applicable in CNN-based tasks without modifying other layers.

  3. Data leveraging: Maximinzing usability from independent data to achieve TTA within constraints, even with small batch sizes.

- **Approach:**

  By fine-tuning the first representation of the imput image, fast adaptation D_t is possible. Therefore, instead of expensive entropy, we extract and minimize the channel-wise uncertainty from the reconstructed stem layer.



Ours

Introduction - 4

- **Goals:**

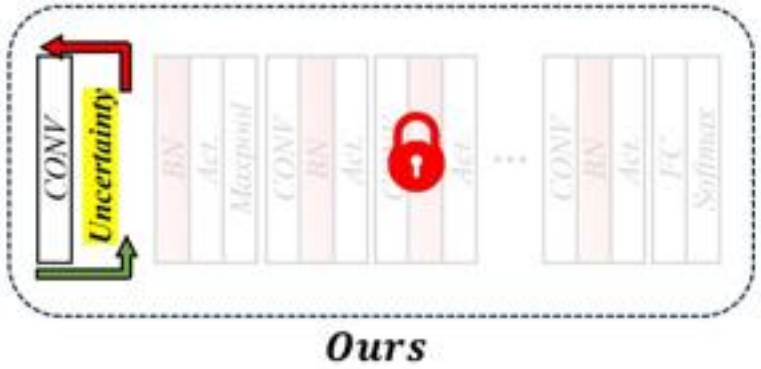
  1. Practicality in training: Minimizing the resources required for training such as memory, **to attain acceptable a reasonable prediction accuracy** in a D_t.

  2. Scalability: Designing to be non-invasively and conveniently applicable in CNN-based tasks without modifying other layers.

  3. Data leveraging: Maximinzing usability from independent data to achieve TTA within constraints, even with small batch sizes.

- **Approach:**

  By fine-tuning the first representation of the imput image, fast adaptation D_t is possible. Therefore, instead of expensive entropy, we extract and minimize the channel-wise uncertainty from the reconstructed stem layer.



Introduction - 4

- **Goals:**

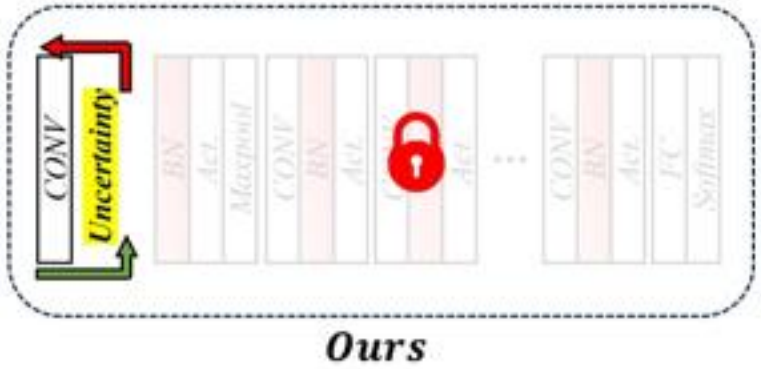
  1. Practicality in training: Minimizing the resources required for training such as memory, **to attain acceptable a reasonable prediction accuracy** in a D_t.

  2. Scalability: Designing to be non-invasively and conveniently applicable in CNN-based tasks without modifying other layers.

  3. Data leveraging: Maximinzing usability from independent data to achieve TTA within constraints, even with small batch sizes.

- **Approach:**

  By fine-tuning the first representation of the imput image, fast adaptation D_t is possible. Therefore, instead of expensive entropy, we extract and minimize the channel-wise uncertainty from the reconstructed stem layer.



Introduction - 4

- **Goals:**

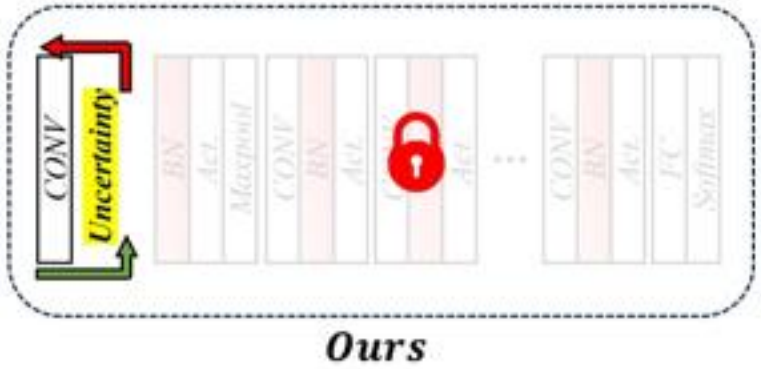
  1. Practicality in training: Minimizing the resources required for training such as memory, **to attain acceptable a reasonable prediction accuracy** in a D_t.

  2. Scalability: Designing to be non-invasively and conveniently applicable in CNN-based tasks without modifying other layers.

  3. Data leveraging: Maximinzing usability from independent data to achieve TTA within constraints, even with small batch sizes.

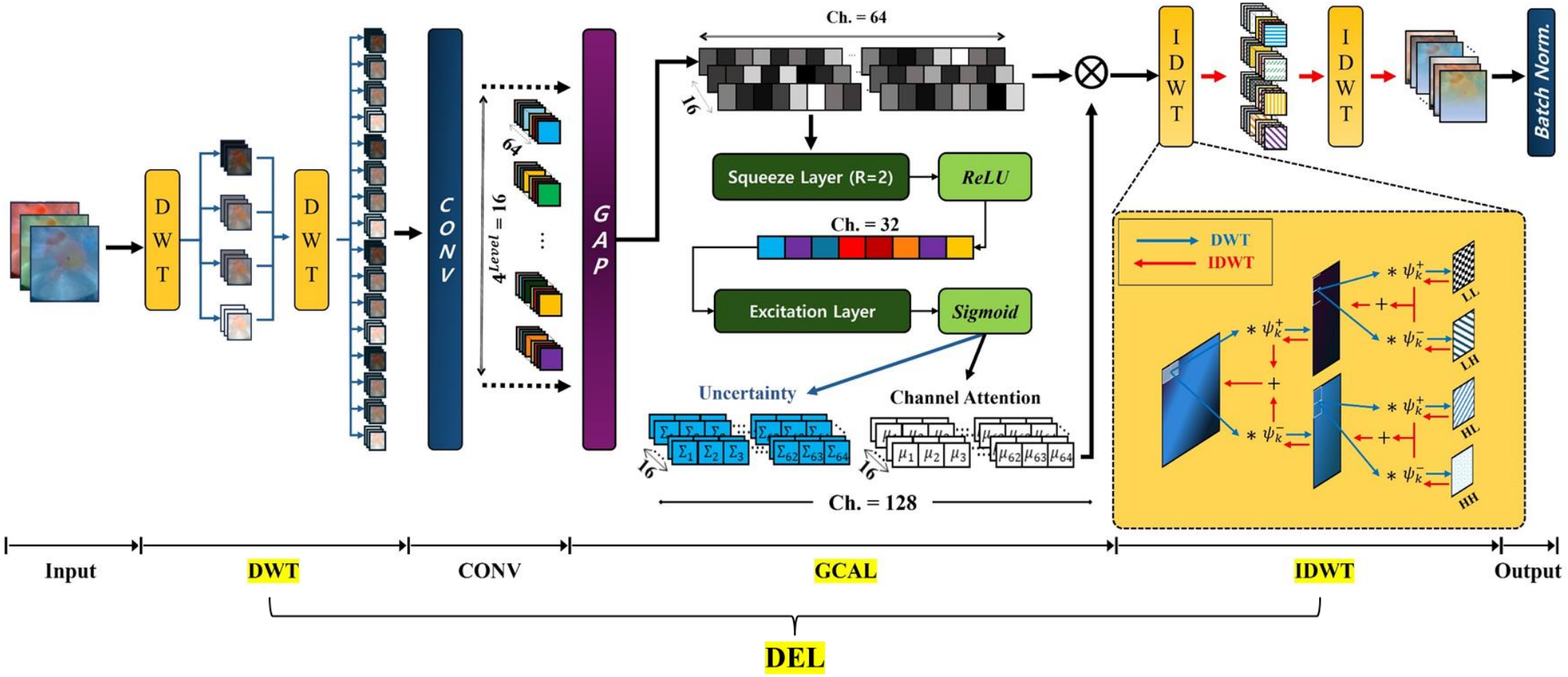
- **Approach:**

  By fine-tuning the first representation of the input image, fast adaptation D_t is possible. Therefore, instead of expensive entropy, we extract and minimize the channel-wise uncertainty from the reconstructed stem layer.



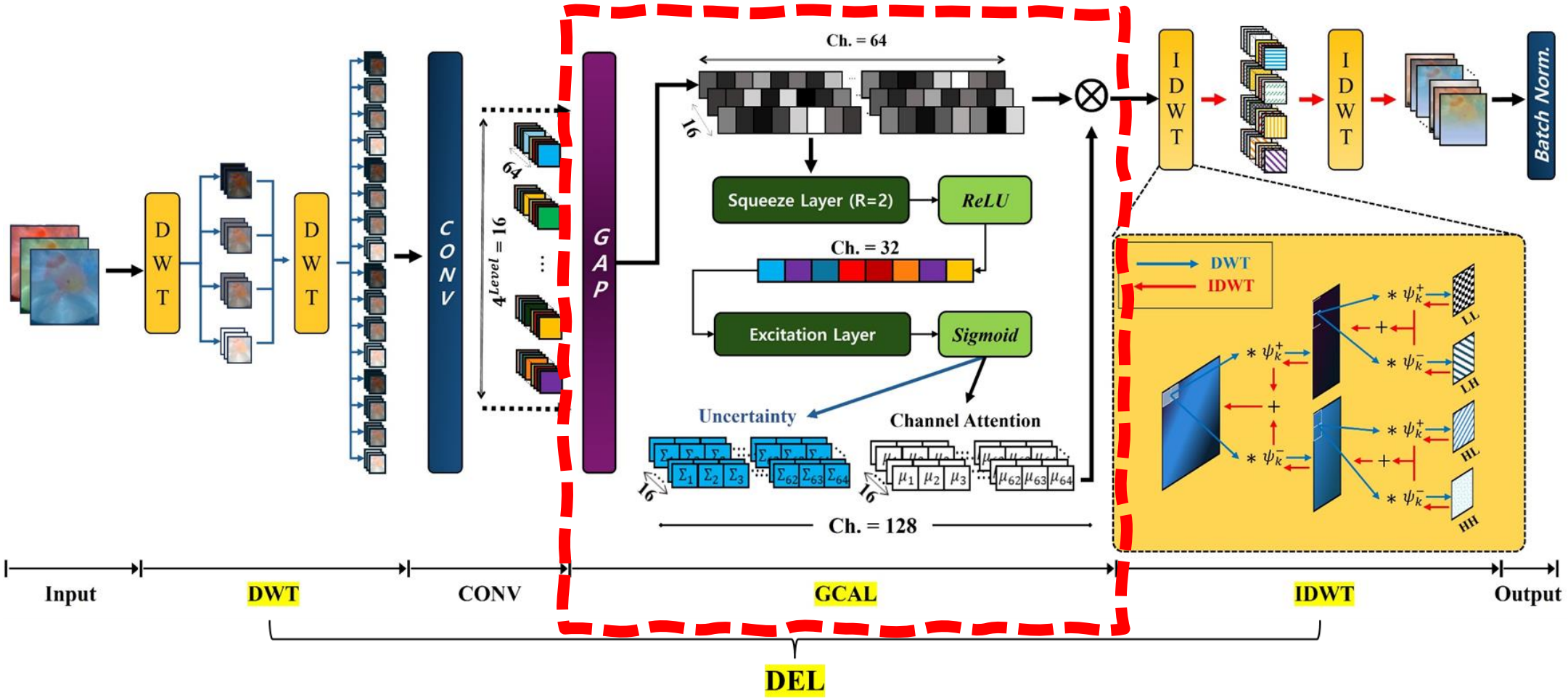
L-TTA - Overview

- Overview of our proposed method including reconstructed stem layer



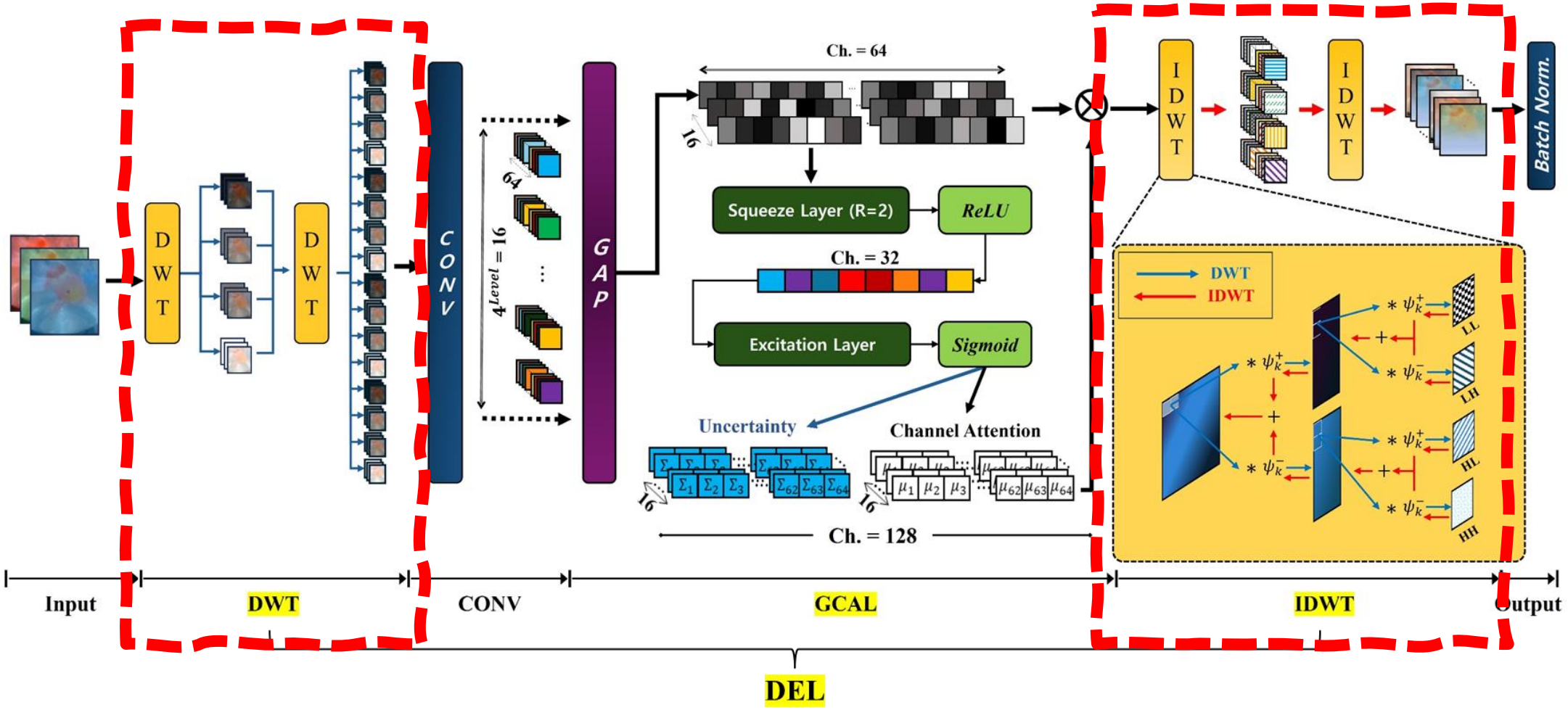
L-TTA - Gaussian Channel Attention Layer

- Overview of our proposed method including reconstructed stem layer



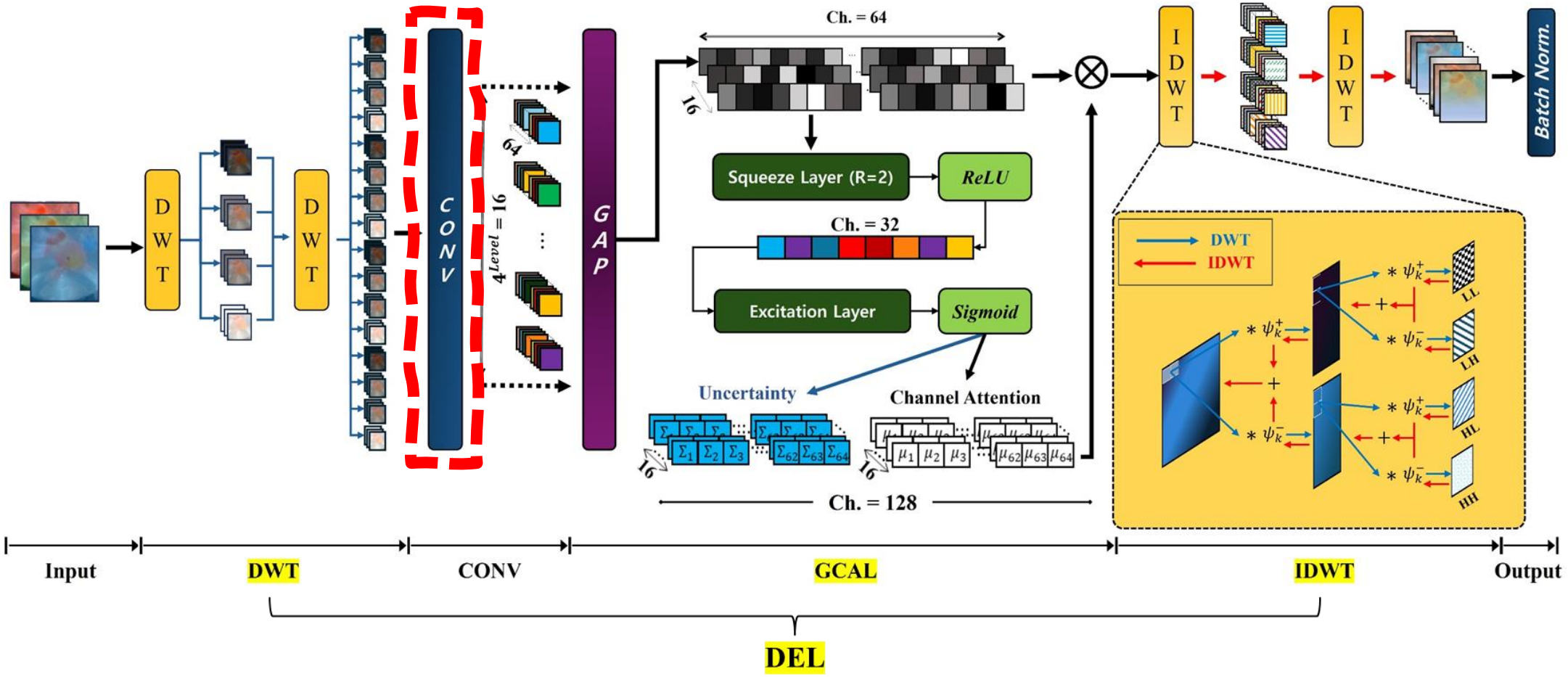
L-TTA - Domain Embedding Layer

- Overview of our proposed method including reconstructed stem layer



L-TTA – Baseline

- Overview of our proposed method including reconstructed stem layer

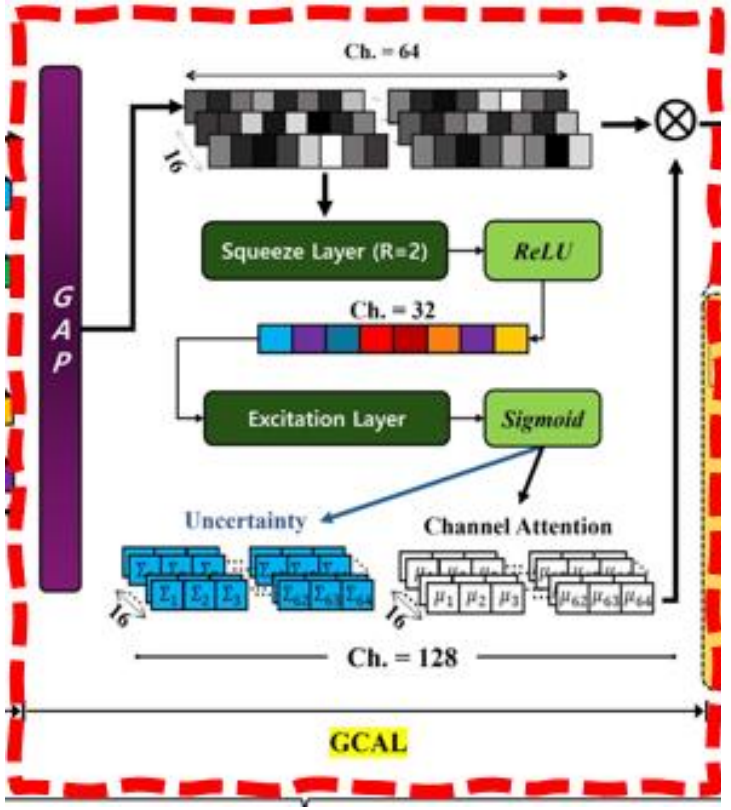


L-TTA - Gaussian Channel Attention Layer

- **GCAL: Gaussian Channel Attention Layer**

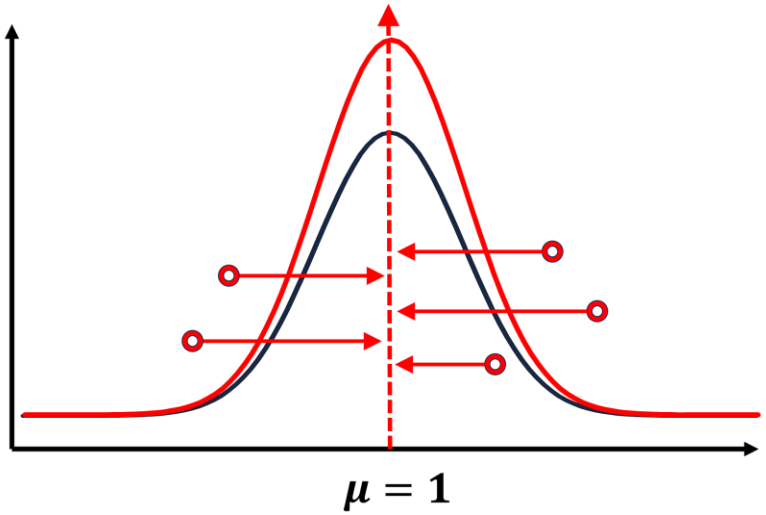
  - The output channels of the squeeze and excitation layer are doubled, with half allocated to channel-wise uncertainty.

  - All channels are encouraged to produce the correct output for inputs with low uncertainty when pretrained on D_s with GCAL, as shown in the equation below:

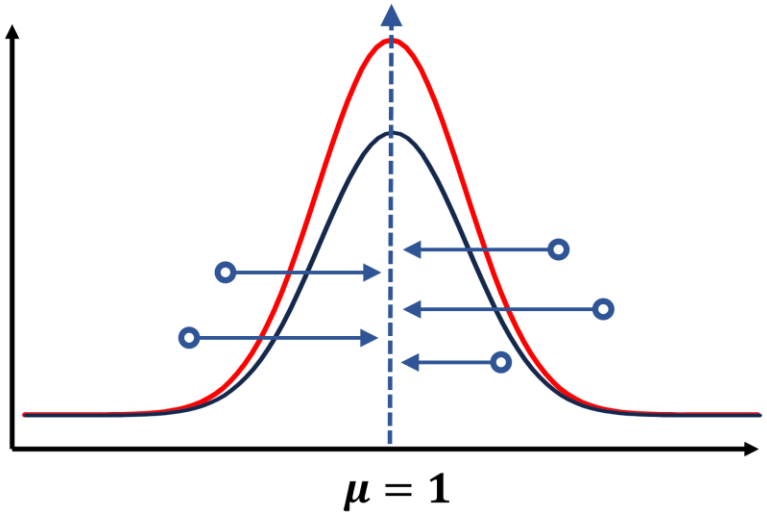


$$L_{uncertainty} = \frac{1}{C}\sum_{i=0}^{C-1} -log(p_i(\mu_{gt}; \gamma_\mu, \gamma_\Sigma)),$$

* Fixed to 1



(a) ***Pretraining on*** D_s



(b) ***TTA processing on*** D_t

L-TTA - Gaussian Channel Attention Layer

- **GCAL: Gaussian Channel Attention Layer**

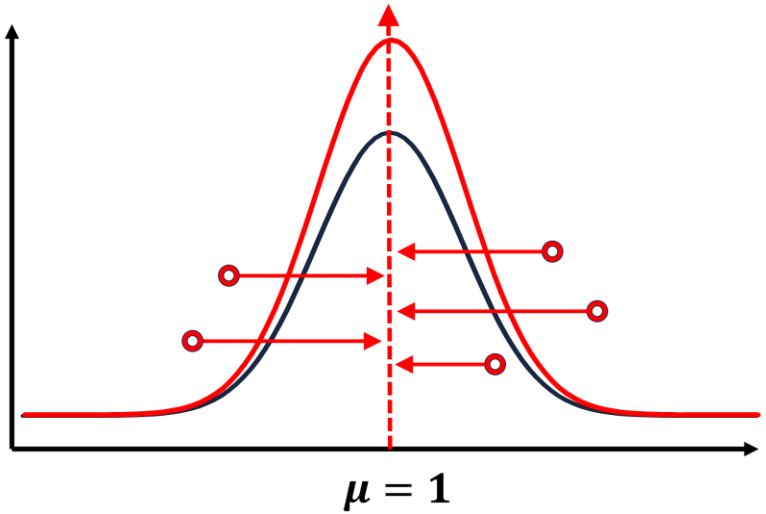
  - Obviously, the D_t has a potentially higher uncertainty due to the different data distribution (especially, in the high-frequency domain)

  - Based on these two facts, we minimize uncertainty about the unlabeled data in the D_t.

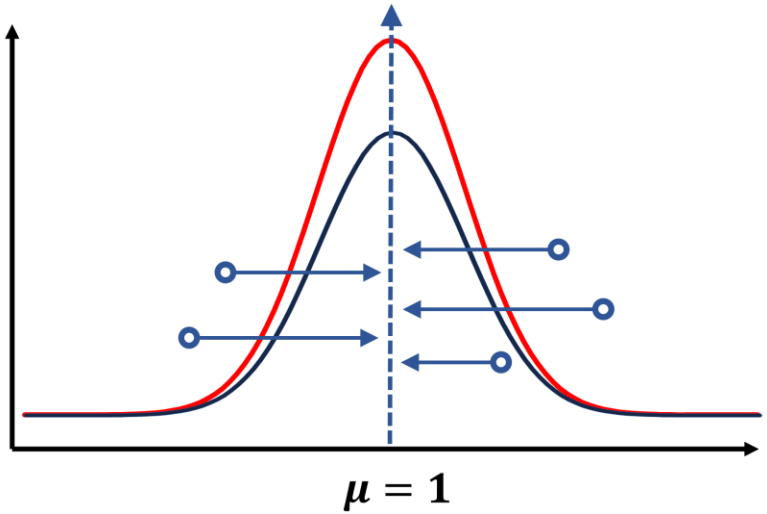
  - This process helps ensure that the performance of frozen subsequent layers can be leveraged like it was in the D_s.

$$L_{uncertainty} = \frac{1}{C}\sum_{i=0}^{C-1} -log(p_i(\mu_{gt}; \gamma_\mu, \gamma_\Sigma)),$$

* Fixed to 1



(a) ***Pretraining on*** D_s



(b) ***TTA processing on*** D_t

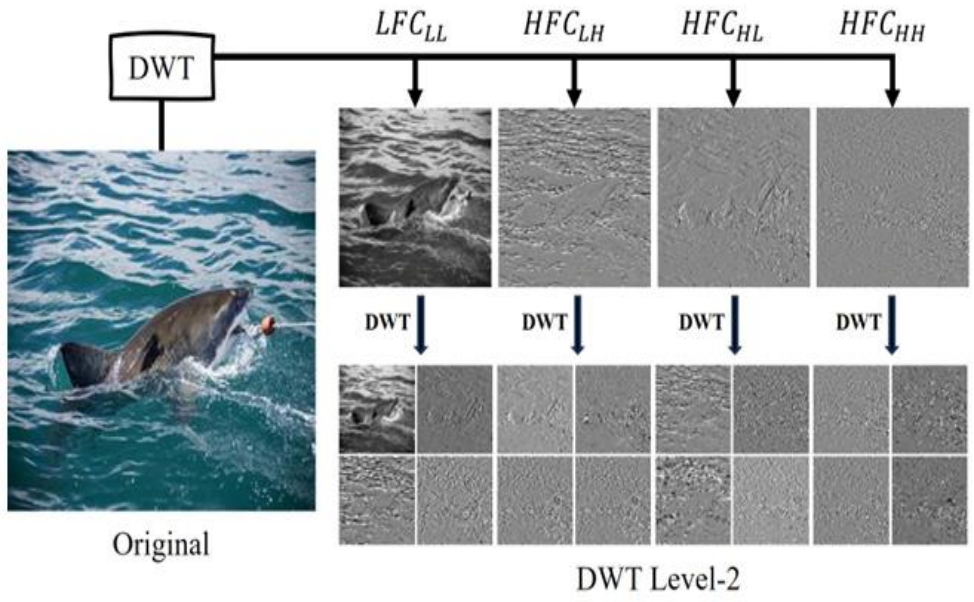
L-TTA - Domain Embedding Layer

- **DEL: Domain Embedding Layer**

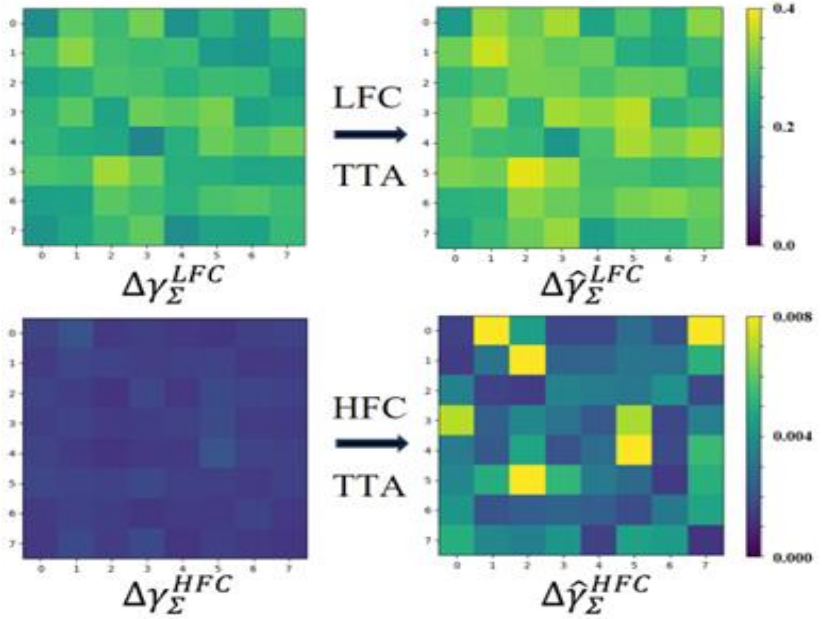
  - To alleviate entropy ambiguity, we propose DEL, which encapsulates GCAL and CONV layers with discrete wavelet transform (DWT) and Inverse DWT layers.

  - By decomposing a single data into low- and high-frequency domains through DWT, we can derive low and high entropy respectively.

  - Our reconstructed stem layer allow us to identify and minimize entropy within these freuquency domains



< **ODD**: **O**mni**D**irectional **D**ecomposition >

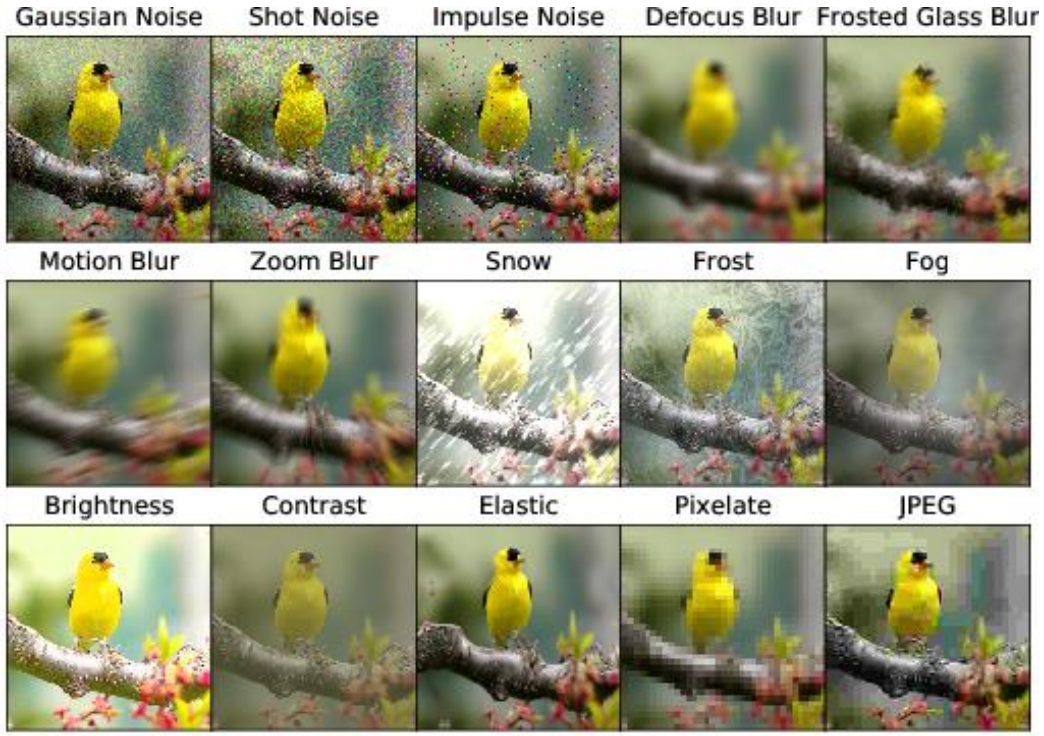


< The variance of uncertainty according to the domain shift of LFC and HFC >

Experimental Results - 1

- **Image Classification (REALM vs. L-TTA)**
  - **Scenario 1: CIFAR-10 → CIFAR-10-C (+ 5.4%)**
  - **Scenario 2: ImageNet → ImageNet-C (– 2.7%)**

Method	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
CIFAR-10-C																
Source	86.9	82.6	81.8	11.4	50.2	18.9	9.1	16.4	26.4	18.4	7.1	24.5	22.8	64.0	28.3	36.6
Ours w/o GCAL	62.4	55.4	48.5	12.1	51.7	15.5	8.5	16.6	26.4	21.7	6.4	31.0	21.3	61.7	22.5	30.8
TENT[59]	39.4	38.8	47.9	19.9	45.0	23.2	20.6	28.1	32.1	24.5	16.1	26.7	32.4	30.6	35.5	30.7
MEMO[65]	43.5	39.9	43.3	26.4	44.4	25.1	25.0	20.9	28.3	22.8	11.9	28.3	21.1	42.8	21.7	29.7
SFT[33]	31.9	26.7	**28.9**	17.7	44.2	18.4	20.2	20.8	23.4	20.7	13.9	25.4	24.5	21.9	25.1	24.2
EATA[45]	33.9	32.8	41.4	19.4	42.4	20.5	20.1	22.4	27.1	22.7	13.8	24.0	24.0	29.3	26.5	26.7
SAR[46]	46.4	40.9	50.1	20.2	47.0	21.7	20.8	22.9	29.5	23.9	13.8	25.5	24.3	39.5	27.4	30.3
REALM[53]	**27.8**	**25.4**	35.5	15.5	37.7	17.4	16.9	20.4	22.3	19.0	12.9	18.0	23.1	22.0	24.2	22.5
Ours w/o DEL	34.1	30.0	32.3	9.9	37.5	11.4	7.3	**12.9**	**13.3**	**11.7**	5.7	**7.0**	18.3	16.2	24.8	18.2
Ours	31.7	27.4	30.9	**9.1**	**35.1**	**11.3**	**6.9**	**12.9**	13.7	12.2	**5.5**	7.7	**18.0**	**12.6**	**22.6**	**17.2**
ImageNet-C																
Source	97.8	97.1	98.2	82.1	90.2	85.2	77.5	83.1	76.7	75.6	41.1	94.6	83.1	79.4	68.4	82.0
Ours w/o GCAL	82.7	82.6	86.2	79.1	89.1	84.2	75.7	74.8	67.3	73.1	37.5	84.6	74.1	43.2	55.2	72.6
TENT[59]	97.5	97.1	97.5	86.5	96.4	81.4	82.4	84.7	77.0	98.6	29.6	57.8	93.8	50.8	46.2	78.5
MEMO[65]	81.5	79.5	81.6	82.9	87.4	78.2	73.1	59.6	**53.0**	65.6	30.5	63.5	80.8	67.9	46.7	68.8
DDA[17]	**57.6**	**56.7**	**57.7**	83.4	**80.4**	78.1	74.0	64.3	59.9	86.3	38.8	74.8	62.5	53.4	45.9	64.9
EATA[45]	75.2	71.7	74.3	81.9	82.7	71.5	70.7	55.5	55.7	58.4	29.1	**55.4**	73.0	53.2	44.3	63.5
SAR[46]	76.6	73.4	76.1	81.6	84.6	71.4	69.6	55.1	55.3	74.3	**27.7**	55.5	85.2	53.0	**43.9**	65.5
REALM[53]	73.1	70.1	72.0	81.6	81.8	**70.4**	68.9	**54.4**	56.4	54.5	28.8	55.6	71.1	50.3	44.5	**62.2**
Ours w/o DEL	75.1	74.5	76.1	84.7	88.3	76.0	68.3	62.3	53.1	58.9	30.5	81.8	64.5	62.6	57.0	67.6
Ours	79.0	78.4	81.8	**75.8**	81.5	72.9	**64.1**	69.1	54.3	**58.8**	33.6	73.3	**57.6**	**41.0**	54.3	65.0

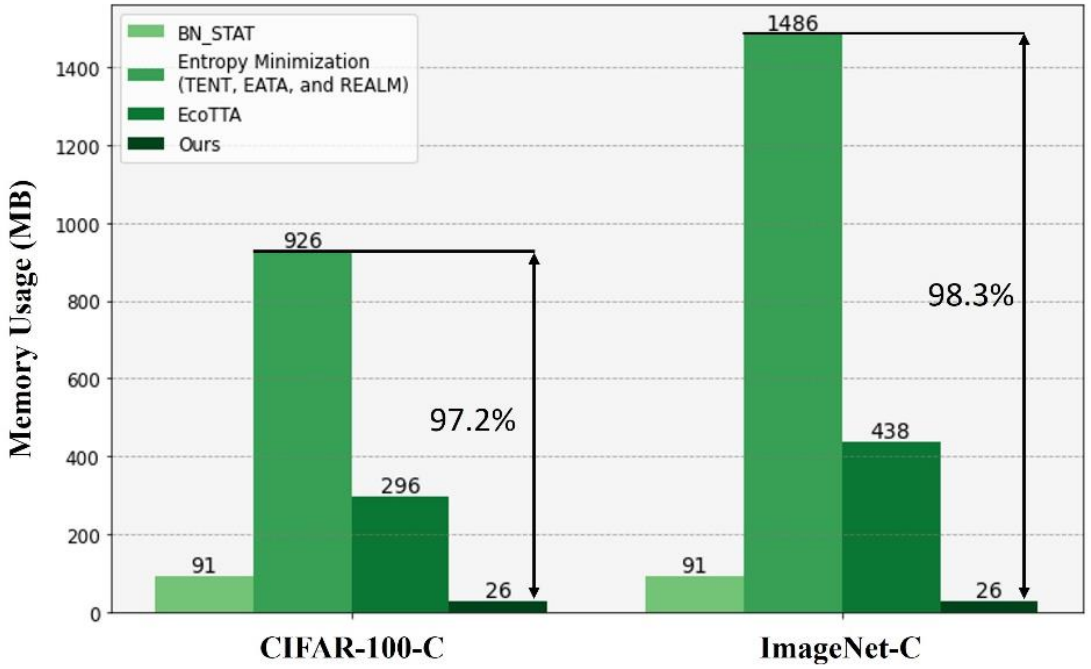


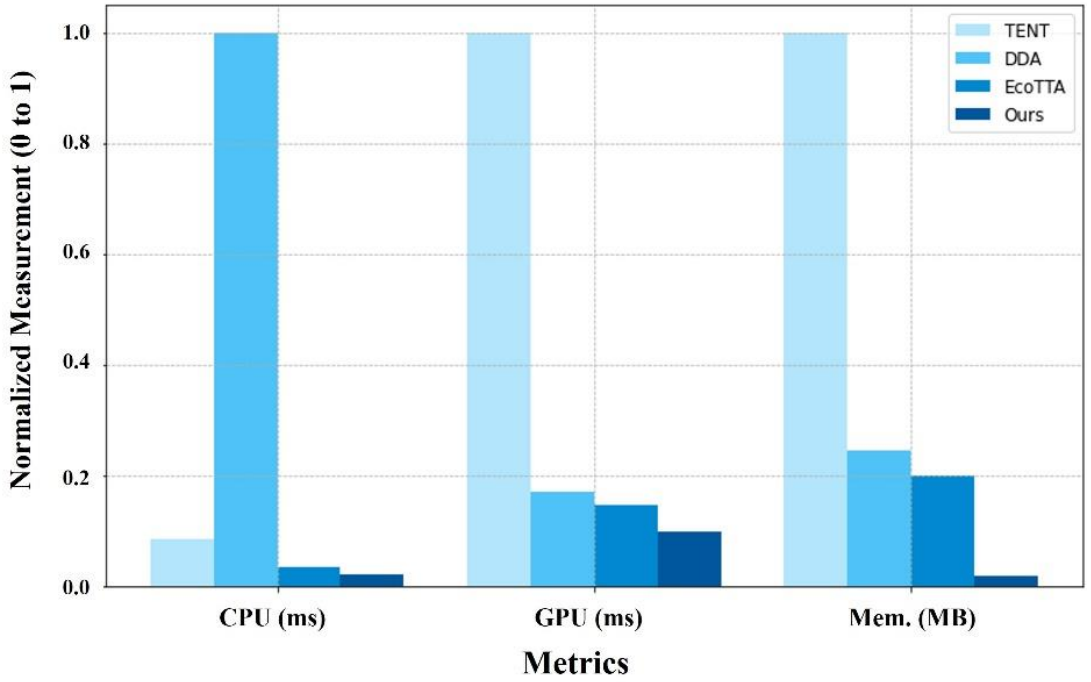
< 15 types of corruptions with most severe level (=5) >

Experimental Results - 2

- **Consumed Memory Usage (EM vs. EcoTTA vs. L-TTA)**
  - **CIFAR-100-C (2.8%, and 8.8%)**
  - **ImageNet-C (1.7%, and 5.9%)**

- **Training time (EM vs. DDA vs. EcoTTA vs. L-TTA)**
  - **L-TTA demonstrates 4.25×, 49.56×, and 1.76× faster in total latency.**





Thank You!