Changing Modalities: Adapting Remote Sensing Models to New Satellites and Sensors

Tim G. Zhou, Anthony Fuller, Geoff Pleiss, Evan Shelhamer

Modalities can change, so models must change

- New satellites launch with novel sensors, old ones retire;
- Re-labeling data for every new modality is costly;

Goal: Adapt existing models to new modalities without any new labels

Three practical scenarios of changing modalities

	Labeled Split	Adaptation Split	Evaluation Split
Transfer	$\mathcal{M}_A, \mathcal{Y}$	$\mathcal{M}_A \& \mathcal{M}_B$	$\mathcal{M}_B$
Addition			$\mathcal{M}_A \& \mathcal{M}_B$
Peeking			$\mathcal{M}_A$

⇔ Transfer—Switch from an old modality to a new one

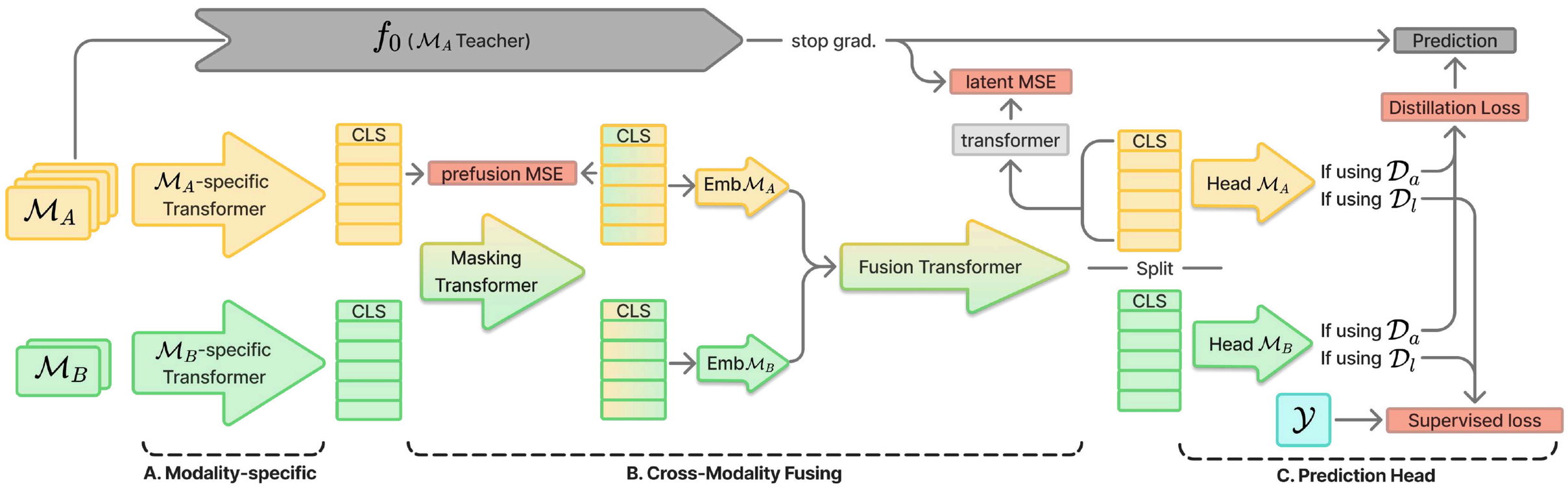
Addition—Predict using both old and new modalities.

Peeking—Learn knowledge from a new modality to improve predictions on the original one

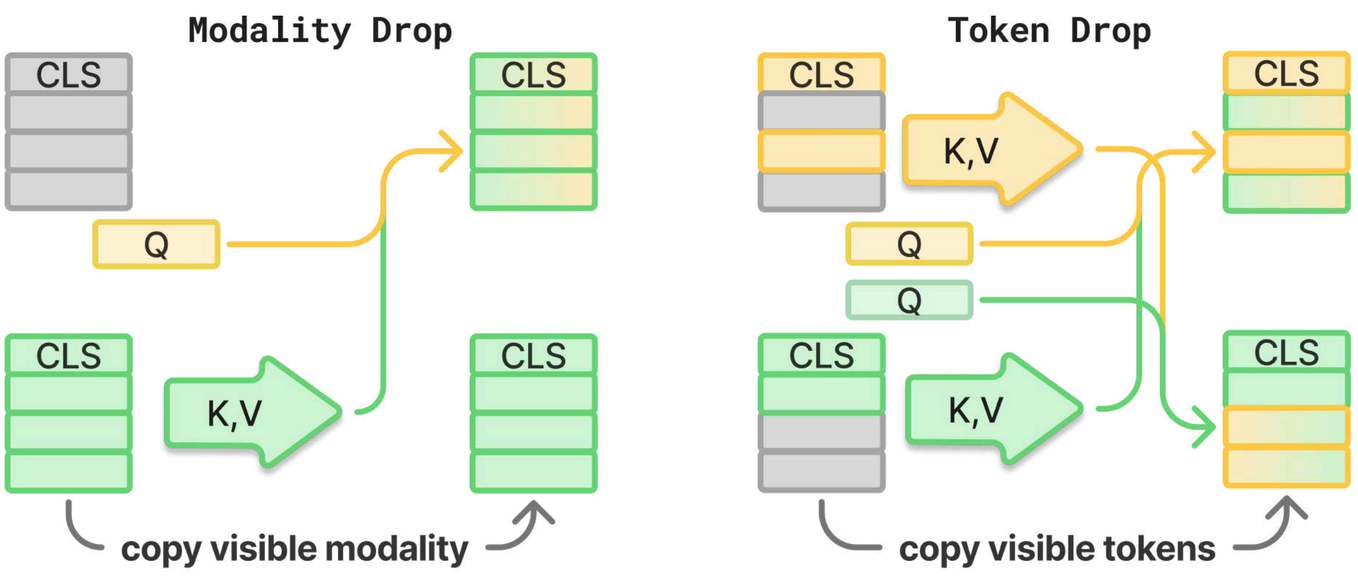
DeluluNet: One method to cover them all via *modality hallucination*

Starts from any uni-modal ViT (f_0) trained with labeled M_A data, adapted into a DeluluNet

1. Learn M_B-specific feature extractor and modality-fusion components with self-supervision.
2. Batch Mixing: Learn to predict using real multimodal input and pseudo labels, or pseudo multimodal input with real labels.
3. Modality-hallucination allows predicting with any subset of modalities, enabling transfer, peeking, and addition together.



Masking Transformers



A closer look at the cross-attention layer in Masking Transformer under different train-time masking strategies.

- At train time, masking transformers facilitate cross-modality fusion using masked image modeling.
- At test time, transfer and peeking scenarios activates modality-drop path for the unavailable modality, and modality addition scenario does not require masking transformers.

$M_A \to M_B$		Transfer reBEN (mAP) S1 → S2	Transfer DFC2020 (mIoU) $S2_{rgb} \to S2_{\neg rgb}$	Addition reBEN (mAP) S1 → S2	Addition DFC2020 (mIoU) $S2_{rgb} \to S2_{\neg rgb}$	Peeking reBEN (mAP) S1 → S2	Peeking DFC2020 (mIoU) $S2_{rgb} \to S2_{\neg rgb}$
$f_0(M_A)$	f_0	61.3	41.9	61.3	41.9	61.3	41.9
Baselines	KD	49.3 ± 0.0	42.4 ± 2.2	–	–	–	–
	TTM	48.7 ± 2.1	44.9 ± 2.2	–	–	–	–
	MKE	–	–	60.8 ± 0.2	39.1 ± 2.7	–	–
	MixMatch	–	–	–	–	59.5 ± 2.3	39.1 ± 3.4
Ours	Delulu	**49.9 ± 0.3**	**50.6 ± 1.4**	**61.4 ± 0.6**	**47.0 ± 2.1**	**62.5 ± 0.6**	**46.7 ± 1.3**
Oracle	DINOv3	49.2	39.9	61.5	49.5	–	–
	Panopticon	55.1	48.5	63.4	50.8	62.1	45.5

⇔ Transfer: DeluluNet significantly improves upon knowledge distillation (KD) and Transformed Teacher Matching (TTM).

Addition: DeluluNet learns and predicts with both modalities, outperforming multimodal knowledge expansion (MKE).

Peeking: DeluluNet learns to make uni-modal predictions with multi-modal observations, outperforming MixMatch.

What's next?

Ever-Changing Modalities*:* Scale from two-modality experiments to sequences of changing sensors, chaining transfer, addition, and peeking as satellites are introduced or retired, stepping towards truly sustainable deployment of RS models.