FUSION-TO-BRANCH DISTILLATION FOR LABEL-EFFICIENT MULTIMODAL ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Label-efficient multimodal learning uses complementary views to reduce the need for labeled data. In multimodal human activity recognition (HAR), this requires a strong complete-view predictor and modality-specific predictors that remain accurate when only one sensor stream is available. Existing semi-supervised methods either create targets from one view or exchange predictions between views, so no branch learns from targets based on both modalities. We introduce fusion-to-branch distillation (F2B), which transfers the complete-view posterior to each modality branch on paired unlabeled examples. A single training run produces one fusion predictor and two single-modality predictors. F2B requires neither a pretrained teacher nor an additional network at inference. We evaluate F2B on five HAR datasets using four label budgets. Compared with supervised training, F2B improves fusion performance in 17 of 20 dataset-budget settings. It improves the direct branches in 38 of 40 evaluations. In 17 of 20 settings, the average branch improvement is also larger than the fusion improvement. Ablations on three datasets show that each branch benefits more from predictions based on the other modality than from its own, supporting cross-modal transfer, although the small number of independent evaluation units limits statistical power. At inference, using one branch requires 44-56% fewer floating-point operations than full fusion. Overall, F2B improves single-modality prediction while retaining the benefits of multimodal fusion in one model.

1 INTRODUCTION

A skeleton tracker can lose the subject to occlusion, a wearable can disconnect, and privacy or power limits can leave only one active sensor. Human activity recognition models pair pose or skeleton with inertial data during training, yet at test time either source can drop out on its own. This creates two coupled problems: labels are scarce, and the model must still predict accurately from each individual source, not only from the two combined.

Synchronized observations provide supervision that labels do not. A *geometric* stream, meaning pose or skeleton, describes global body configuration, while an inertial measurement unit (IMU) captures local acceleration and rotation. We use "geometric" throughout because the datasets supply pose in some cases and skeleton joints in others. Even without an activity label, the two recordings describe the same event. Conventional semi-supervised learning derives targets from one prediction stream (Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2023). Multimodal pretraining can align paired streams but requires a later fine-tuning stage for each target modality (Ouyang et al., 2022; Deldari et al., 2024; Xu et al., 2023). Neither design trains the fused predictor to teach each direct branch within the same training run.

Our main finding is that paired unlabeled data improve the direct branches more consistently than fusion, and that a branch learns more from a teacher built on the other sensor than from its own prediction. We study this through *fusion-to-branch distillation* (F2B), summarized in Figure 1. Two modality encoders feed branch classifiers and a fusion head. On each paired unlabeled example, the detached fusion prediction becomes a soft target for both branches under strong augmentation. Because this target is computed from both sensors, it gives each branch evidence that its own input does not contain. The teacher is a head in the same model, so F2B needs no pretrained teacher and adds no computation at test time.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

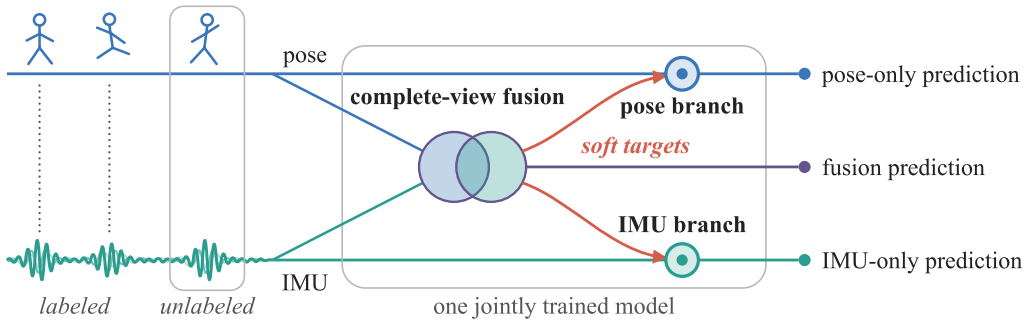


Figure 1: Overview of F2B. During training, the fusion head combines pose or skeleton with IMU and gives each direct branch a soft target on paired unlabeled data, shown as red arrows. At test time, the same model predicts from both inputs or from either input alone.

The evaluation separates three questions that are easily conflated: whether paired unlabeled data improve fusion, whether they improve the direct branches, and whether the fusion head still works when one input is set to zero. It covers five pose or skeleton and IMU datasets with segmented and continuous recordings, subject-disjoint evaluation, and four label fractions, with teacher-source controls on three datasets and unimodal specialist comparisons on two.

Our contributions are:

- fusion-to-branch distillation, a joint-training objective in which the fusion head teaches each modality branch on paired unlabeled data, with no pretrained teacher and with both multimodal and single-sensor prediction retained,
- a teacher-source decomposition on three datasets showing that an other-branch teacher outperforms a same-branch teacher at both direct branches, which supports cross-view transfer beyond self-training, and
- a five-dataset, four-budget evaluation showing that the direct branches gain more consistently than fusion, and that F2B branches perform on par with semi-supervised unimodal specialists while keeping a fusion output from one training run.

2 Related Work

Consistency-based semi-supervised methods turn a confident weak-view prediction into supervision for a stronger view (Tarvainen & Valpola, 2017; Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2023). Co-training instead exchanges predictions between informative views (Blum & Mitchell, 1998; Qiao et al., 2018), and multi-head co-training and cross pseudo supervision exchange targets between predictors of the same input (Chen et al., 2022b; 2021a). F2B differs in what teacher and student observe: the complete-view predictor teaches a branch that cannot access all of its evidence. The transfer is therefore hierarchical rather than reciprocal.

F2B follows online knowledge distillation, where teacher and students are trained together. Deep mutual learning uses symmetric peer teaching (Zhang et al., 2018), while ONE distills a gated ensemble into its constituent branches using a stopped target (Lan et al., 2018). F2B instead assigns branches to different modalities, so the fusion teacher combines distinct sensory evidence. Prior complete-view-to-unimodal methods pretrain a labeled teacher and target a particular missing-input configuration (Chen et al., 2022a; Wei et al., 2023). Multimodal knowledge expansion transfers in the opposite direction, from a pretrained unimodal teacher to a multimodal student (Xue et al., 2021). F2B jointly trains the fusion teacher and both direct branches from labeled and paired unlabeled data. Its gate is a standard gated multimodal unit (Arevalo et al., 2017), not a contribution.

Cross-modal methods align or transfer information between synchronized observations (Baltrusaitis et al., 2019). CoSMo learns a contrastive fused representation, CroSSL masks latent time-series features, and MESEN pretrains multimodally before fine-tuning one unimodal model (Ouyang et al., 2022; Deldari et al., 2024; Xu et al., 2023). Other methods distill between skeleton modalities or

transfer privileged video evidence to an action or IMU model (Mao et al., 2022; Garcia et al., 2018; Ni et al., 2022; Chen et al., 2025), and DMCL selects a low-loss labeled modality as teacher (Garcia et al., 2021). F2B uses paired unlabeled observations and retains fusion and every direct branch after one joint training run.

Joint multimodal training can favor one source and leave weak unimodal representations, a behavior described as modality laziness or imbalance (Wang et al., 2020; Peng et al., 2022; Du et al., 2023). Incomplete-modality methods train with masked inputs (M3L), align or translate missing representations (CMAT), learn from missing modalities and labels (MUSE), or adapt the model at test time (MiDl) (Maheshwari et al., 2024; Park et al., 2023; Wu et al., 2024; Ramazanova et al., 2025). X-Fi learns modality-invariant representations, MoRA adapts pretrained vision-language models, and DyMo selects recovered modalities at inference (Chen & Yang, 2025; Zhao et al., 2026; Du et al., 2026). F2B studies a different setting: modalities are paired during training, and complete-view predictions on unlabeled data supervise each direct branch without reconstruction or test-time adaptation.

3 Fusion-to-branch distillation

3.1 Problem and model

Let $\mathcal{D}_L = \{(x_g^i, x_a^i, y^i)\}_{i=1}^{N_L}$ contain labeled synchronized geometric (pose or skeleton) and inertial inputs, and let $\mathcal{D}_U = \{(x_g^j, x_a^j)\}_{j=1}^{N_U}$ be a disjoint pool of paired unlabeled inputs. Encoders f_g, f_a produce representations h_g, h_a, and a linear classifier on each representation gives the branch logits $z_g = \phi_g(h_g)$ and $z_a = \phi_a(h_a)$. The fusion head combines them in two parts. A gate $\alpha = \mathrm{softmax}(\phi_\alpha([h_g \,\|\, h_a]))$ weights the two branch logits for each sample, and a residual head ϕ_r adds a correction computed from both representations:

$$z_f = \alpha_g z_g + \alpha_a z_a + \phi_r([h_g \,\|\, h_a]), \qquad p_m = \mathrm{softmax}(z_m),\ m \in \{g, a, f\}. \tag{1}$$

The model therefore has three outputs: the complete-view output p_f, which uses both inputs, and two *direct branches*, p_g and p_a, each of which needs only its own encoder. Figure 2 shows the full training graph, and the labeled objective supervises all three outputs:

$$\mathcal{L}_{\mathrm{sup}} = \mathrm{CE}(z_f, y) + \frac{\lambda_{\mathrm{aux}}}{2}\left[\mathrm{CE}(z_g, y) + \mathrm{CE}(z_a, y)\right]. \tag{2}$$

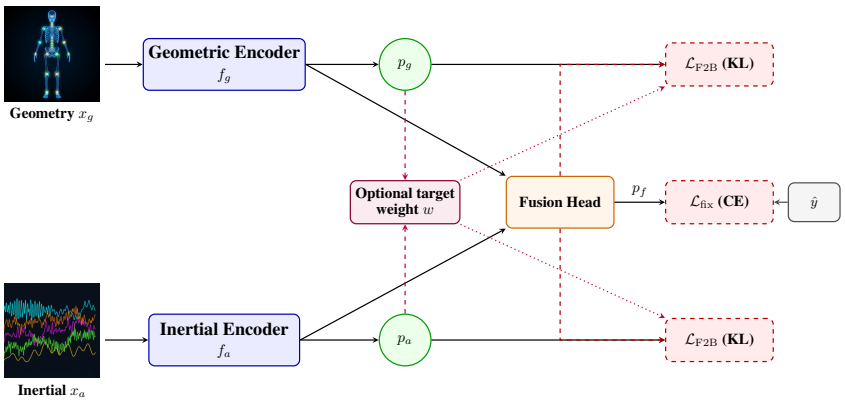


Figure 2: F2B training on one unlabeled pair. The weak-view fusion posterior p_f is detached and used as the soft target for both strongly augmented branches, shown as red dashed arrows, when the pair passes the class-adaptive threshold. Dashed boxes are training-only losses. The purple optional weight w belongs to RACM. Base F2B sets $w = b$.

3.2 Distillation on paired unlabeled data

For each unlabeled pair j, the model sees a weakly and a strongly augmented view, with logits z_m^j and $\widetilde{z}_m^j$ for output m. On the weak view, the fusion head predicts a pseudo-label $\hat{y}^j = \arg\max_c p_f^j(c)$

with confidence $q_f^j = \max_c p_f^j(c)$. The pair is used only when this confidence passes a class-adaptive threshold, similar to FreeMatch (Wang et al., 2023):

$$b^j = \mathbf{1}[q_f^j \geq \tau_{\hat{y}^j}], \qquad \tau_c = \begin{cases} \tau_0, & n_c < k_{\min}, \\ \bar{q}_c, & n_c \geq k_{\min}. \end{cases} \tag{3}$$

Here n_c counts the pseudo-labels assigned to class c, and $\bar{q}_c$ is a moving average of their confidence. A class therefore starts from the fixed threshold τ_0 and switches to its own threshold once it has $k_{\min}$ pseudo-labels.

Each pair enters two unlabeled losses with weight w^j, where $w^j = b^j$ in base F2B. The first, $\mathcal{L}_{\text{fix}}$, trains the strong-view fusion logits on the pseudo-label with cross-entropy, as in FixMatch (Sohn et al., 2020). The second is the F2B loss. It uses the weak-view fusion posterior as a soft target for both strong-view branches:

$$\mathcal{L}_{\text{F2B}} = \frac{1}{\sum_j w^j} \sum_j \frac{w^j}{2} \sum_{m \in \{g,a\}} \tau_s^2 \, \text{KL}\Big(\text{sg}[\text{softmax}(z_f^j/\tau_t)] \,\|\, \text{softmax}(\tilde{z}_m^j/\tau_s)\Big). \tag{4}$$

Here $\text{sg}[\cdot]$ denotes stop-gradient. We use teacher temperature $\tau_t = 1$ and student temperature $\tau_s = 2$, fixed before evaluation. The factor τ_s^2 keeps the gradient size comparable across student temperatures. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \gamma_e \left(\lambda_{\text{fix}} \mathcal{L}_{\text{fix}} + \lambda_{\text{F2B}} \mathcal{L}_{\text{F2B}}\right), \tag{5}$$

where γ_e is zero during a supervised warm-up and then rises linearly to one. F2B changes only training and adds no computation at test time.

3.3 What each branch learns

For a fixed teacher and an unrestricted branch predictor, minimizing the expected distillation loss gives

$$\pi_m^*(\cdot \mid x_m) = \arg\min_{\pi} \mathbb{E}[\text{KL}(p_f(\cdot \mid x_g, x_a) \,\|\, \pi(\cdot \mid x_m)) \mid x_m] = \mathbb{E}[p_f(\cdot \mid x_g, x_a) \mid x_m]. \tag{6}$$

Each branch therefore learns the fusion prediction averaged over all paired observations that are consistent with its own input. If the fusion head matched the true posterior $p(y \mid x_g, x_a)$, this target would equal the best single-modality posterior $p(y \mid x_m)$. F2B thus uses unlabeled pairs to give each branch a soft estimate of that posterior, formed with evidence from both modalities rather than from the branch's own prediction. This links F2B to distillation with privileged information (Lopez-Paz et al., 2016; Garcia et al., 2018) and motivates two design choices. The teacher is the fusion head, because it is the only output that sees both modalities, and a same-branch teacher would only reinforce the branch's own prediction. The target is detached, so the fusion head is not pulled toward the weaker single-sensor predictions. The analysis is idealized, because filtering, augmentation, shared parameters, and the moving teacher constrain the actual optimization.

3.4 Reliability-weighted variant

Reliability-Aware Cross-Modal Consistency (RACM) is a variant that changes only which unlabeled pairs count and how much. It replaces the weight b^j with $w^j = b^j \, r^j \max(w_A^j, w_F^j)$. The indicator r^j keeps a pair if the fusion head is confident, if the two branches agree and one of them is confident, or if both branches agree with the pseudo-label. The soft weight w_A^j grows with the agreement between the two branch predictions, and w_F^j grows with how far the fusion confidence exceeds its threshold. All weights are computed without gradients. RACM keeps both unlabeled losses of F2B unchanged, so comparing RACM with F2B isolates the effect of reliability weighting.

4 Experimental design

4.1 Datasets, labels, and partitions

Table 1 lists five public datasets that pair skeleton or pose with inertial data. UTD-MHAD uses its standard cross-subject split (Chen et al., 2015). CZU-MHAD (Chao et al., 2022) uses five subject-disjoint folds. UTD Kinect V2 uses six leave-one-subject-out folds (Chen et al., 2016). mRI uses

five held-out-subject blocks (An et al., 2022). Continuous MM-Fit uses the official workout split and non-overlapping five-second training windows (Strömbäck et al., 2020). Label fractions are selected within class from the training pool, and remaining training samples form the disjoint paired unlabeled pool. Separate labeled validation partitions select checkpoints and are shared across matched methods. For MM-Fit, the 10% condition uses about 473 labeled windows. No validation example enters a training loss, and test labels never affect training, filtering, or model selection.

Table 1: Paired datasets and evaluation protocols. Samples are segmented trials, except for MM-Fit, which counts non-overlapping five-second windows. All segmented protocols are subject-disjoint, with UTD-MHAD using a single fixed cross-subject partition.

Dataset	Samples	Subjects	Classes	Protocol
UTD-MHAD	861	8	27	fixed cross-subject
CZU-MHAD	1,165	5	22	5 subject folds
UTD Kinect V2	300	6	10	6 LOSO folds
mRI[†]	200	20	10	5 subject blocks
MM-Fit	11,474	21	11	official workout split

[†]mRI selects labels by whole subject, so its four budgets are 7.1/14.3/28.6/50.0%, not 10/15/25/50%.

These benchmarks range from 200 to 11,474 samples, and at the lowest label fraction UTD Kinect V2 and mRI supply only tens of labeled sequences. We therefore base our conclusions on consistency across datasets and budgets rather than any single dataset-budget setting.

4.2 Models and controls

All methods share one base model: a three-block temporal convolutional network (TCN) per modality, 256-dimensional pooled representations, branch classifiers, and the gated residual fusion head. A mechanism control swaps the pose TCN for CTR-GCN (Chen et al., 2021b). Matched methods also share the AdamW optimizer, schedules, augmentations, label subsets, and splits, with three seeds per split, five for MM-Fit.

On every dataset and label budget, we compare supervision, fusion-level FixMatch (Sohn et al., 2020), F2B, and RACM. The MM-Fit ablation study adds (i) *no distillation*, identical to F2B except $\lambda_{\mathrm{F2B}} = 0$, and (ii) *branchwise FixMatch*, sharing architecture, data, augmentations, optimizer, schedule, and unlabeled update but forming a hard target and fixed mask from each branch's own weak prediction. No distillation differs from fusion-level FixMatch by keeping the class-adaptive threshold, and is the stronger of the two on MM-Fit at 10% labels. Adapted CoSMo (Ouyang et al., 2022), DMCL (Garcia et al., 2021), and MESEN (Xu et al., 2023) share encoders and data protocols with F2B. MESEN produces one separately fine-tuned model per target modality and has no fusion output.

We decompose the F2B teacher on MM-Fit at 10% labels with five matched seeds, comparing five teacher sources for each target branch: (i) the *full fusion* logit, (ii) the *other*-branch logit, which tests cross-view transfer, (iii) the *same*-branch logit, which is self-distillation, (iv) *fusion without self*, which drops the target branch's weighted logit but keeps the residual head that sees both inputs, and (v) the *residual head* alone. All variants keep the same masks, augmentations, optimizer, schedule, and F2B loss weight.

The central full-fusion, no-distillation, other-branch, and same-branch contrasts are repeated on mRI and CZU-MHAD with five held-out units. On these datasets, each direct branch is compared with supervised and FixMatch unimodal specialists trained on the same labeled subset. The FixMatch specialists use the same unlabeled pool as F2B.

A separate temperature study compares the default $(\tau_s, \lambda_{\mathrm{F2B}}) = (2, 0.05)$ with $(1, 0.05)$ and $(2, 0.025)$ while fixing $\tau_t = 1$. The last two settings are first-order gradient-scale matched because the F2B gradient carries a factor $\lambda_{\mathrm{F2B}}\tau_s$. This design separates the effect of changing the softened student distribution from the leading change in gradient magnitude. All teacher-source and temperature comparisons average seeds within each of the five unseen workouts before computing paired differences and confidence intervals.

4.3 METRICS AND STATISTICAL TESTING

We report complete-view fusion and the two *direct* branches separately. Segmented datasets use top-1 accuracy. Imbalanced continuous MM-Fit uses macro-F1. A secondary diagnostic sets one input to zero and passes it through the unchanged fusion head. Seeds are averaged within each held-out participant or block before statistical inference, leaving five CZU, six UTD Kinect V2, five mRI, and five MM-Fit units. UTD-MHAD's one fixed test split supports repeatability but not inference across subjects. The prespecified primary test is a paired t-test on participant-level differences. For the direct-branch results, Holm correction is applied within each dataset across its eight budget-by-branch comparisons. Sign-flip values and all other confidence intervals are unadjusted unless a Holm-adjusted p is stated. With five units the sign-flip test cannot fall below 0.0625, so it is read mainly for agreement in direction.

5 RESULTS

5.1 FUSION PERFORMANCE

Across five datasets and four label budgets, F2B improves on supervised fusion in 17 of 20 conditions. It leads the semi-supervised methods in nine conditions, compared with eight for RACM and three for FixMatch. The three decreases are at most 1.4 points, and the clearer pattern appears at the direct branches.

Table 2: Complete-view performance across five datasets and four training-label budgets. Segmented datasets report top-1 accuracy (%) and MM-Fit reports macro-F1. Results average three seeds per split, or five for MM-Fit. Bold marks the highest semi-supervised mean in each row.

Dataset	Labels	Supervised	FixMatch	F2B	RACM
UTD-MHAD	10%	42.56	**47.36**	46.67	45.81
	15%	59.69	60.39	**62.25**	61.63
	25%	72.25	69.38	71.24	**72.48**
	50%	80.16	81.01	82.33	**82.56**
CZU-MHAD	10%	76.61	76.23	79.19	**79.77**
	15%	86.41	88.65	**91.61**	89.42
	25%	90.49	92.05	**93.78**	93.23
	50%	93.16	93.67	**94.34**	93.81
UTD Kinect V2	10%	55.44	59.33	**62.22**	61.00
	15%	69.22	68.33	67.78	**70.33**
	25%	68.78	72.78	**73.00**	72.33
	50%	77.22	**81.22**	78.67	76.89
mRI[†]	10%	71.17	**73.67**	71.50	71.83
	15%	84.17	84.67	86.00	**87.50**
	25%	89.67	92.00	90.33	**94.00**
	50%	95.17	96.33	**96.50**	96.17
MM-Fit	10%	85.86	85.54	**86.76**	85.48
	15%	86.46	85.99	87.34	**87.89**
	25%	87.16	87.99	**90.24**	89.94
	50%	88.47	88.07	87.68	**88.23**

[†] mRI realizes 7.1/14.3/28.6/50.0% labels.

5.2 DIRECT BRANCH PERFORMANCE

Table 3 shows that 38 of 40 branch gains over matched supervision are positive. All 32 segmented-dataset gains are positive, averaging 4.03 pp for the geometric branch and 6.70 pp for the inertial branch. MM-Fit adds six positive gains and two small geometric decreases at the two largest budgets. This comparison evaluates full F2B training against matched supervision, so it shows repeatability rather than the causal contribution of the F2B loss.

The participant-level tests have only five or six units. Ten of 32 unadjusted intervals exclude zero, but only the $+2.4$-pp CZU geometric gain at 50% labels survives Holm correction. We therefore treat these results as consistent but descriptive evidence across correlated conditions.

5.3 ABLATION STUDIES

Table 4 reports the objective and encoder ablations on MM-Fit. F2B exceeds branchwise FixMatch on fusion and on both branches. Because this comparison changes the target, mask, and filtering together, it does not isolate the F2B loss.

Table 3: Direct branches under limited training labels. A direct branch predicts from its own modality alone, without the other encoder or the fusion head. Segmented datasets report top-1 accuracy (%) and MM-Fit reports macro-F1. Parentheses show F2B minus matched supervision in percentage points.

Dataset	Branch	10%	15%	25%	50%
UTD-MHAD	Geometric	37.4 (+1.9)	48.8 (+4.9)	53.8 (+0.9)	60.7 (+1.6)
	Inertial	33.7 (+5.4)	58.1 (+22.0)	69.1 (+4.7)	77.5 (+4.7)
CZU-MHAD	Geometric	73.7 (+5.4)	88.2 (+5.4)	90.1 (+3.5)	91.5 (+2.4)
	Inertial	67.9 (+7.5)	82.9 (+6.6)	84.6 (+5.0)	85.5 (+3.3)
UTD Kinect V2	Geometric	43.2 (+6.6)	44.7 (+1.1)	46.1 (+5.1)	56.3 (+11.6)
	Inertial	60.3 (+2.2)	68.1 (+4.8)	65.3 (+3.6)	77.0 (+6.3)
mRI[†]	Geometric	70.5 (+3.7)	82.7 (+3.5)	90.0 (+5.3)	93.2 (+1.7)
	Inertial	58.5 (+5.8)	79.8 (+12.8)	84.5 (+9.7)	85.8 (+2.7)
MM-Fit	Geometric	71.9 (+1.3)	73.5 (+4.6)	75.2 (−0.3)	74.7 (−0.7)
	Inertial	82.8 (+7.0)	79.5 (+4.6)	82.9 (+1.2)	79.9 (+2.2)

[†]mRI's whole-subject selection realizes 7.1/14.3/28.6/50.0% labels in these columns.

With the TCN encoder, the F2B loss has an asymmetric effect: it improves the inertial branch, while the geometric branch and fusion change little. The inertial direction is stable across the tested temperatures, but the geometric direction is not. We retain the prespecified default rather than select the strongest post hoc result.

Table 4: MM-Fit ablations at 10% labels. Values are macro-F1 over five unseen workouts after seed averaging. The CTR-GCN control uses unit temperature and is not directly comparable with the TCN panel. MESEN has no fusion output. Bold marks the highest joint-model mean within each panel.

Method	Fusion	Geometric	Inertial
TCN geometric encoder			
Supervised	85.86 ± 8.82	70.57 ± 9.42	75.87 ± 9.16
Branchwise FixMatch	83.92 ± 10.41	69.83 ± 10.61	77.96 ± 10.58
No distillation	$\mathbf{86.96 \pm 9.36}$	$\mathbf{72.38 \pm 9.47}$	80.00 ± 8.32
F2B (full fusion teacher)	86.76 ± 8.89	71.88 ± 8.86	$\mathbf{82.83 \pm 10.30}$
MESEN	–	76.60 ± 8.70	83.72 ± 9.28
CTR-GCN geometric encoder			
Supervised	83.13 ± 12.28	55.48 ± 14.59	74.42 ± 7.87
No distillation	$\mathbf{84.28 \pm 12.78}$	59.89 ± 15.63	79.27 ± 8.43
F2B	84.06 ± 13.08	$\mathbf{62.34 \pm 16.18}$	$\mathbf{80.64 \pm 8.54}$

5.4 Teacher-source ablation

Table 5 reports the core teacher-source contrasts on three datasets. The F2B loss raises the inertial branch on every dataset. On CZU-MHAD, fusion and both branches improve with intervals excluding zero, though the geometric branch changes little on MM-Fit and mRI. The clearer test compares teachers directly. An other-branch teacher exceeds a same-branch teacher at both direct branches on all three datasets. Three of the six comparisons have intervals excluding zero. This supports a cross-view contribution beyond self-training.

No reduced teacher dominates across datasets. Removing the student's own logit or using only the residual helps some outputs and harms others. The best teacher composition is therefore branch- and dataset-dependent.

In the weight-matched control, unit temperature exceeds the gradient-matched $\tau_s = 2$ setting for fusion and both branches, showing that temperature is not merely scalar rescaling. Relative to the default, it improves the geometric branch by 1.50 pp and reduces the inertial branch by 1.38 pp. Temperature therefore controls a branch trade-off rather than a common optimum.

Table 5: Teacher-source ablation at 10% labels. Entries are percentage-point differences with unadjusted 95% confidence intervals over five evaluation units. Panel (a) is the full fusion teacher minus no distillation. Panel (b) is the other-branch teacher minus the same-branch teacher. Bold marks intervals excluding zero.

Dataset	Fusion	Geometric	Inertial
(a) F2B loss			
MM-Fit	-0.20 [-1.46, 1.07]	-0.51 [-2.08, 1.06]	$+2.83$ [-0.03, 5.69]
mRI	$+0.33$ [-2.65, 3.31]	-0.67 [-3.03, 1.69]	$+2.50$ [-2.18, 7.18]
CZU-MHAD	$+\mathbf{3.34}$ [0.70, 5.98]	$+\mathbf{3.71}$ [0.40, 7.02]	$+\mathbf{6.48}$ [4.67, 8.30]
(b) Teacher source			
MM-Fit	$+1.74$ [-0.49, 3.97]	$+\mathbf{4.45}$ [0.22, 8.68]	$+\mathbf{2.24}$ [0.60, 3.88]
mRI	-0.33 [-2.48, 1.81]	$+1.83$ [-0.84, 4.51]	$+3.33$ [-1.57, 8.24]
CZU-MHAD	$+\mathbf{3.59}$ [0.65, 6.54]	$+\mathbf{3.82}$ [1.45, 6.20]	$+5.38$ [-0.65, 11.41]

5.5 Unimodal specialist comparison

A jointly trained branch can be weaker than a model trained on that modality alone (Wang et al., 2020; Peng et al., 2022). Table 6 therefore compares each branch with a single-encoder specialist using the same labeled subset and unlabeled pool.

The direct branches exceed supervised specialists in all four comparisons. This argues against the branch gains being mere recovery from weak joint training. Against FixMatch specialists that also use unlabeled data, all intervals include zero. MESEN's stronger MM-Fit geometric branch supports the same conclusion: F2B is a versatile single model, not a uniformly stronger unimodal learner.

Table 6: Direct branches relative to unimodal specialists at 10% labels. Entries are percentage-point differences with unadjusted 95% confidence intervals over five subject folds. FixMatch specialists use the same unlabeled pool. Bold marks intervals excluding zero.

Dataset	Branch	vs. supervised specialist	vs. FixMatch specialist
mRI	Geometric	$+1.50$ [-2.33, 5.33]	-1.67 [-6.46, 3.13]
	Inertial	$+\mathbf{7.33}$ [1.55, 13.11]	$+0.50$ [-4.27, 5.27]
CZU-MHAD	Geometric	$+1.02$ [-2.08, 4.12]	-1.00 [-5.84, 3.84]
	Inertial	$+2.73$ [-0.30, 5.75]	$+1.29$ [-1.32, 3.90]

5.6 Robustness and baseline comparisons

RACM and F2B reach similar fusion performance, RACM leading eight of the 20 conditions and F2B nine. RACM is more consistent in the zero-input stress test, improving 31 of 32 segmented conditions by 8.7 pp on average. Against adapted CoSMo and DMCL, RACM is better in 21 of 24 matched comparisons, equal in one, and worse in two. The average gain is 8.25 pp, though none of these differences survives family-wise correction. RACM is therefore a robustness option when the fusion head must accept a zeroed input, while the direct branches remain the recommended predictors when only one sensor is available.

6 Discussion

6.1 Main findings

The gains from paired unlabeled data concentrate at the direct branches. As Figure 3 shows, the mean branch gain exceeds the fusion gain in most conditions, and all three conditions with a lower fusion score still have a positive mean branch gain. F2B therefore improves the single-sensor predictors without a systematic loss at fusion. Fusion and branch results answer different questions. Fusion measures prediction with all inputs, while a direct branch measures whether one sensor can work alone. One trained model provides all three outputs, and using a direct branch reduces approximate inference cost from 212.1M FLOPs to 118.1M or 93.4M.

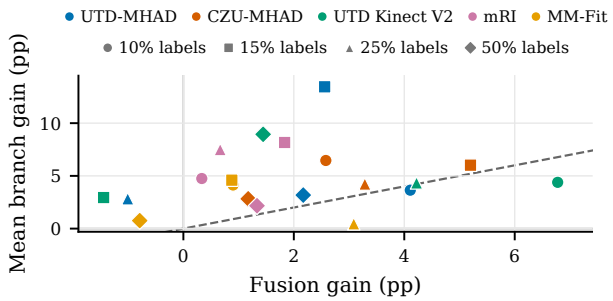


Figure 3: Distribution of F2B gains across 20 dataset-budget conditions. The horizontal axis shows the fusion change and the vertical axis the mean direct-branch change. Seventeen points lie above equality. Segmented datasets use accuracy and MM-Fit uses macro-F1, so only within-condition differences are compared.

6.2 Mechanism

The ablations support the cross-view account of what each branch learns. On all three datasets, an other-branch teacher exceeds a same-branch teacher at both direct branches, while removing the self or residual components helps some outputs and harms others. F2B is therefore better described as a learned mixture of cross-view, self, and residual evidence than as ordinary self-training. Temperature shifts this benefit between the branches, which is consistent with modality imbalance in joint training (Wang et al., 2020; Peng et al., 2022; Du et al., 2023) and suggests tuning it for the branch needed at test time.

The specialist comparisons qualify the accuracy claim. F2B branches exceed supervised specialists on mRI and CZU-MHAD but are statistically indistinguishable from FixMatch specialists that use the same unlabeled pool, and MESEN is stronger for the MM-Fit geometric branch. The practical advantage is therefore one model that supports changing sensor availability, not uniformly higher unimodal accuracy.

6.3 Limitations

The ablation studies cover fewer settings than the main evaluation. Teacher-source controls cover three datasets at 10% labels, specialist controls cover two, and the temperature study covers only MM-Fit. Statistical power is also limited. Four datasets provide only five or six held-out units, UTD-MHAD uses one fixed partition, and only one direct-branch gain survives Holm correction. Larger subject cohorts are needed for tighter estimates.

The study covers synchronized pose or skeleton and inertial data, assumes known sensor availability at test time, and does not test temporal misalignment, corrupted sensors, more than two modalities, or other domains. Label fractions exclude the validation partition, and cost measurements use one GPU at batch size one.

7 Conclusion

Fusion-to-branch distillation uses paired unlabeled data to improve the direct branches of a multimodal model while keeping its fusion output. Across five datasets and four label budgets, the direct branches improve in 38 of 40 settings and gain more than fusion in 17 of 20. Teacher-source controls support a cross-view contribution beyond self-training, and specialist comparisons show that the main benefit is flexible inference from one trained model rather than higher unimodal accuracy than dedicated semi-supervised models. These results suggest that multimodal learning should also evaluate the branches needed when a sensor is unavailable, not only the predictor that uses every sensor.

AI USE STATEMENT

In this work, we have not used generative AI tools for any task with required disclosure under the ICLR 2027 AI Policy for Authors. We used generative AI tools for tasks with recommended disclosure: identifying relevant literature and searching for information, and drafting and editing parts of the paper to improve readability. We have reviewed all AI-assisted work. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses only publicly available datasets that were collected and released by their original authors, and it does not involve any new data collection from human subjects. Because activity recognition from body pose and wearable sensors concerns personal data, such methods should be applied only with appropriate consent and privacy safeguards.

REPRODUCIBILITY STATEMENT

Section 4 describes the datasets, subject-disjoint splits, label budgets, and evaluation protocol. The supplementary material gives the full training protocol, hyperparameters, additional controls, and statistical tests. All datasets used in this work are publicly available, and code and experiment configurations will be released after the review period.

REFERENCES

Sizhe An, Yin Li, and Umit Ogras. mRI: Multi-modal 3d human pose estimation dataset using mmWave, RGB-D, and inertial sensors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27414–27426, 2022.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *International Conference on Learning Representations (Workshop Track)*, 2017.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pp. 92–100, 1998.

Xin Chao, Zhenjie Hou, and Yujian Mo. CZU-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and 10 wearable inertial sensors. *arXiv preprint arXiv:2202.03283*, 2022. doi: 10.48550/arXiv.2202.03283.

Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora D. Salim. COMODO: Cross-modal video-to-imu distillation for efficient egocentric human activity recognition. *arXiv preprint arXiv:2503.07259*, 2025. doi: 10.48550/arXiv.2503.07259.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Fusion of depth, skeleton, and inertial data for human action recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2712–2716, 2016. doi: 10.1109/ICASSP.2016.7472170.

Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng-Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Transactions on Medical Imaging*, 41(3): 621–632, 2022a.

Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. Semi-supervised learning with multi-head co-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6278–6286, 2022b. doi: 10.1609/aaai.v36i6.20577.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021a.

Xinyan Chen and Jianfei Yang. X-Fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, 2025.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13359–13368, 2021b.

Shohreh Deldari, Dimitris Spathis, Mohammad Malekzadeh, Fahim Kawsar, Flora D. Salim, and Akhil Mathur. CroSSL: Cross-modal self-supervised learning for time-series through latent masking. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 152–160, 2024. doi: 10.1145/3616855.3635795.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8632–8656, 2023.

Siyi Du, Xinzhe Luo, Declan O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *International Conference on Learning Representations*, 2026.

Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision*, pp. 103–118, 2018.

Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2755–2764, 2021.

Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1020–1030, 2024.

Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. CMD: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European Conference on Computer Vision*, 2022.

Jianyuan Ni, Anne H. H. Ngu, and Yan Yan. Progressive cross-modal knowledge distillation for human action recognition. *arXiv preprint arXiv:2208.08090*, 2022. doi: 10.48550/arXiv.2208.08090.

Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*, pp. 324–337, 2022. doi: 10.1145/3495243.3560519.

Yeonju Park, Sangmin Woo, Sumin Lee, Muhammad Adi Nugroho, and Changick Kim. Cross-modal alignment and translation for missing modality action recognition. *Computer Vision and Image Understanding*, 236:103805, 2023. doi: 10.1016/j.cviu.2023.103805.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. *arXiv preprint arXiv:1803.05984*, 2018. doi: 10.48550/arXiv.1803.05984.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *International Conference on Learning Representations*, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 596–608, 2020.

David Strömbäck, Sangxia Huang, and Valentin Radu. MM-Fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–22, 2020. doi: 10.1145/3432231.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1195–1204, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: Self-adaptive thresholding for semi-supervised learning. International Conference on Learning Representations (ICLR), 2023. URL `https://openreview.net/forum?id=PDrUPTXJI_A`.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, 2024.

Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, and Jiming Chen. MESEN: Exploit multimodal data to design unimodal human activity recognition with few labels. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*, pp. 1–14, 2023. doi: 10.1145/3625687.3625782.

Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18408–18419, 2021.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. MoRA: Missing modality low-rank adaptation for visual recognition. In *International Conference on Learning Representations*, 2026.