

CAST: Channel-Aware Spatial Transfer Learning with Pseudo-Image Radar for Sign Language Recognition

Md. Shakhoyat Rahman Shujon¹, Sheikh Md. Galib Mahim¹, Md. Milon Islam¹, Md Rezwanul Haque²,
Md Rabiul Islam³, Hamdi Altaheri⁴, Fakhri Karray^{2,5}

¹Department of Computer Science and Engineering, Khulna University of Engineering & Technology

²Department of Electrical and Computer Engineering, University of Waterloo

³Department of Electrical and Computer Engineering, Texas A&M University

⁴College of Applied Computer Science, King Saud University

⁵Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence

¹skt104.shujon@gmail.com, ¹galibmahim01@gmail.com, ¹milonislam@cse.kuet.ac.bd,

²rezwan@uwaterloo.ca, ³rabiul.islam@tamu.edu, ⁴altaheri@ksu.edu.sa, ^{2,5}karray@uwaterloo.ca

Abstract

We propose CAST, a dual-stream architecture that utilizes channel-aware spatial transfer learning for isolated sign language recognition addressing the challenges of magnitude-only 60 GHz radar Range-Time Maps (RTM). The proposed framework combines three physics-aware architectures with pretrained vision backbones, which operate under radar-only constraints across clinical and alphabetical gestures. First, an explicit decibel-to-linear inversion is combined with a windowed fast Fourier transform that extracts Cadence Velocity Diagrams (CVD) while avoiding the harmonic artifacts that arise from the spectral analysis of log-compressed signals. Second, a cross-antenna spatial attention module applies attention to raw antenna channels before the convolution, preserving inter-receiver amplitude covariance. Third, an asymmetric cross-attention mechanism fuses representations from parallel ConvNeXt-Tiny (CVD) and EfficientNetV2-S (RTM) backbones. Extensive experiments reveal that the architecture achieves a Top-1 accuracy of 80.5% under 5-fold cross-validation, establishing a 3.3% improvement over the best single-model baseline (77.2%). The findings suggest that physics-aware signal representations form a promising direction for radar-only sign language recognition under constrained sensor modalities. The source code is available at: <https://github.com/Shakhoyat/CAST-at-SignEval2026>.

1. Introduction

Vision-based sign language recognition has achieved relatively good performance in controlled laboratory environ-

ments, with recent multimodal systems reporting over 99% accuracy on the Italian sign language vocabulary of the MultiMeDaLIS benchmark [9]. In vision-based systems, the most challenging issue is that performance degrades significantly when the camera is removed. Clinical environments illustrate this problem clearly. Hospitals routinely handle patients who are deaf or hard-of-hearing, but using cameras for patients raises regulation and consent issues that are difficult to resolve [14]. Millimeter-wave radar provides a privacy-preserving solution. It captures the kinematics of hand and arm motion while inherently anonymizing the signer by operating at 60 GHz. Hence, it is appropriate for privacy-sensitive applications where continuous visual capture is difficult.

The CVPR 2026 MSLR workshop challenge (Track 2) turns this scenario into a benchmark [4, 17]. The MultiMeDaLIS dataset [1, 14] provides 126 Italian sign language classes (100 medical terms and 26 alphabet letters) captured with an Infineon BGT60TR13C 60 GHz FMCW radar across 205 sessions. In this setup, only Range-Time Maps (RTMs) are available for evaluation; no video, depth, or complex-valued radar data are used.

To address these constraints, the proposed approach in this challenge converts RTMs into three-channel arrays and applies pretrained 2-D Convolutional Neural Networks (CNNs). The key point is that it removes two physical properties of the RTM that are important for recognition. The physical properties are as follows:

1. **Velocity blindness:** The temporal axis in RTM encodes chronological kinematics to represent motion over time, not space. When this is treated as a static image, the temporal dimension is reduced, and important information, including velocity and periodicity, disappears, which

helps to distinguish similar signs.

2. **Antenna-geometry ignorance:** The radar module places three receive antennas in an L-shape (two azimuth, one elevation). Using them as RGB channels (RX1, RX2, and RX3) ignores the geometric information and spacing between antennas.

Contributions: The contribution of this work is to focus on physics-aware architectures of existing methods, designed for magnitude-only radar data. The proposed architecture, Channel-Aware Spatial Transfer (CAST) learning, addresses the above gaps through three modules:

- **CVD extraction with dB-to-linear inversion:** We generate a Cadence Velocity Diagram (CVD) from magnitude-only RTMs by inverting the dB to a linear scale. Then, we apply a Blackman-Harris window Fast Fourier Transform (FFT) along the temporal axis. The linearization is necessary because applying a Fourier transform to log-scale data generates harmonic artifacts with no physical interpretation. This is due to the logarithm transformation violating the linear superposition assumption of Fourier analysis.
- **Cross-Antenna Spatial Attention (CASA):** CASA encodes each antenna independently rather than treating three antenna channels as RGB-like inputs, stacks them into a sequence, and applies multi-head self-attention across antenna positions. The module processes raw antenna signals before the first convolution, preserving inter-receiver amplitude covariance.
- **Asymmetric cross-attention fusion:** The RTM stream is utilized as the query, and the CVD stream provides keys and values. This allows the model to selectively retrieve velocity information when the structural range is insufficient for recognition.

A controlled evaluation protocol is used to isolate the contribution of each architectural module. Under the same single-model setting, CAST outperforms the baseline by 3.3% (80.5% vs. 77.2%). The difference between single-model and overall evaluation is discussed later. Furthermore, we analyze the practical limitations caused by the 13 fps capture rate and failure modes due to sensor physics.

2. Related Work

Radar-based gesture and sign language recognition: Fine-grained gesture sensing with 60 GHz mmWave Frequency Modulated Continuous Wave (FMCW) radar was established by Project Soli [11], which demonstrated that micro-scale hand movements generate distinctive micro-Doppler signatures at millimetre wavelengths. Early radar gesture frameworks converted micro-Doppler spectrograms or Range-Doppler Maps (RDMs) to image tensors and applied 2-D CNNs pretrained on natural images [23]. However, this introduces a representational mismatch as Im-

ageNet kernels capture visual textures rather than phase-based frequency modulations in radar signals. More recent radar-based architectures address this by employing hybrid CNN-LSTM networks that separate spatial convolution from temporal sequence modeling [25], and with Multi-view De-interference Transformers that separate gestures from background noise [8].

For sign language, the TRACE architecture [16] is the most closely related prior work. TRACE employs a residual autoencoder to reduce 128×1024 RDMs to a 256-D bottleneck, followed by a six-layer, eight-head Transformer classifier, achieving 93.6% accuracy on the same 126-class vocabulary. A related text-aligned variant [15] further exploits language supervision for radar representations. However, both works use full complex-valued Range-Doppler data at significantly higher resolutions and frame rates than the RTMs available in our challenge. Therefore, a direct accuracy comparison is not meaningful. The authors of [6, 9] demonstrated that late fusion of radar and video logits can degrade accuracy when the modalities have large gaps, confirming that radar-only recognition requires a customized architecture rather than generic fusion methods.

To our knowledge, no directly comparable RTM-only baseline for 60 GHz radar sign language recognition has been published outside the MSLR workshop series. Existing magnitude-only radar gesture studies are limited to vocabularies of 16 or fewer classes, and all systems reporting above 96% accuracy rely on multi-modal fusion with Doppler or angle-of-arrival information. Prior radar-only sign language work assumes access to complex-valued Range-Doppler data at higher frame rates.

Pseudo-Doppler extraction from magnitude data: Standard Doppler extraction requires complex-valued radar data. When only magnitude is available, temporal spectral analysis of the range-profile time series can still recover the cadence of periodic motions [10]. In literature, FFT-based analysis of magnitude signals is already used in gait and activity recognition to extract micro-Doppler-like patterns from amplitude observations [10]. However, our contribution is more specific: we demonstrate that applying the FFT directly to dB-compressed data introduces harmonic artifacts with no physical interpretation, and that performing a dB-to-linear inversion prior to FFT is a minimal correction required to preserve physically interpretable cadence information.

To our knowledge, this first-order approximation step has not been validated in radar sign language recognition, where the input is restricted to magnitude-based RTMs rather than full RDMs. The kinematic envelope of RDMs carries dynamic signatures that complement static spatial features [20]. In the proposed architecture, this information is not available; hence, accurate spectral extraction from magnitude data becomes the key design issue compared to post-hoc feature selection.

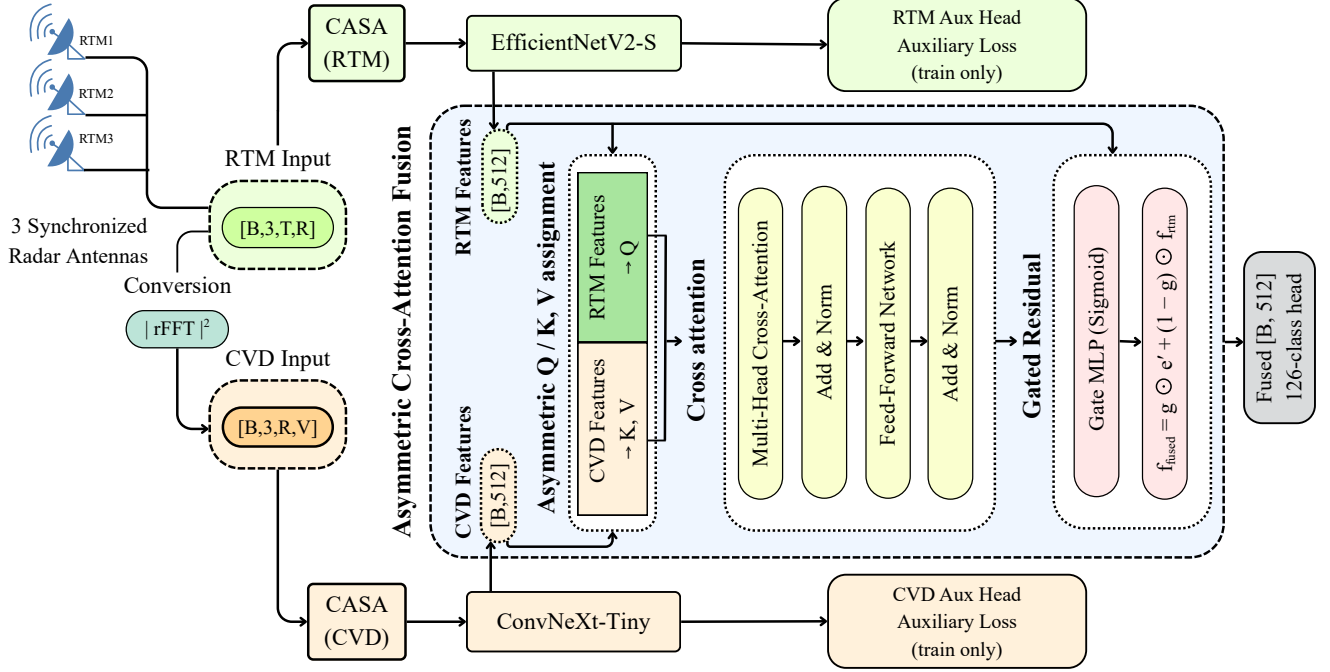


Figure 1. Overall architecture of the proposed CAST architecture. Three-receiver RTMs are processed through per-stream CASA modules and fed into EfficientNetV2-S (RTM) and ConvNeXt-Tiny (CVD) backbones. An asymmetric cross-attention fusion module fuses the dual representations, with a gated residual ensuring fallback to RTM-only features when the CVD stream provides no meaningful information.

3. Method

The overall architecture of our proposed system is illustrated in Fig. 1. There are two parallel branches in this framework: a top stream designed for normalized magnitude RTMs and a bottom stream that extracts CVD. Both streams pass the multi-receiver radar sequences through a CASA module and are then passed through their respective EfficientNetV2-S and ConvNeXt-Tiny backbones for feature extraction. RTM and CVD data are filtered and combined using a cross-attention fusion module, in which the RTM stream selectively retrieves additional information from the CVD streams. This fused information is then used for final recognition.

3.1. Cadence Velocity Diagram Extraction

The MultiMeDaLIS RTMs are distributed as float32 arrays, $\mathbf{R}_{\text{dB}} \in \mathbb{R}^{T \times 256}$ representing $20 \log_{10}(\text{amplitude})$, where T is the number of slow-time frames (≈ 20 – 40 at 13 fps) and 256 is the number of positive range bins (indexed by r).

Necessity of decibel-to-linear inversion: A Fourier transform applied directly to logarithmic dB data is mathematically inconsistent. The logarithm converts a multiplicative modulation of the carrier by a sinusoidal Doppler envelope into an additive form, which violates the linearity assumption required for valid Fourier analysis. An ideal sinusoidal modulation of amplitude $A(1 + m \cos(2\pi f_0 t))$ becomes $\log A +$

$\log(1 + m \cos(2\pi f_0 t))$ after log compression. This transformation does not generate a single spectral peak at f_0 , it generates an infinite series of harmonics, distorting the frequency axis. Expanding $\log(1 + m \cos(2\pi f_0 t))$ as a Taylor series for $|m| < 1$ generates $m \cos(2\pi f_0 t) - \frac{m^2}{2} \cos^2(2\pi f_0 t) + \dots$, which contains terms at $2f_0$, $3f_0$, and all higher harmonics. These higher-order terms generate artificial frequencies that do not correspond to the signer’s actual physical kinematics. This effect is not only theoretical; in practice, applying the FFT directly to dB-scale RTMs creates false peaks at integer multiples of the true cadence frequency, leading the model to learn physically invalid features. Prior magnitude-based cadence analysis in other domains mostly operates on linear amplitude or on datasets where the log compression is limited. The full dB dynamic range available in the MultiMeDaLIS RTMs (>40 dB) amplifies the artifacts, resulting in a 1.7% drop in accuracy (Table 2). The first-order approximation recovers the original amplitude modulation, as shown in (1).

$$\mathbf{R}_{\text{lin}} = 10^{\mathbf{R}_{\text{dB}}/20} \quad (1)$$

Windowed FFT: A Blackman-Harris window $w[n]$ is applied along the temporal axis. Blackman-Harris achieves >92 dB sidelobe suppression, which is essential because dominant torso reflections (broad peaks in range) can mask weaker cadence signals created by hand motion. The window

length T (zero-padded to $N_{\text{FFT}} = 128$), and the positive-frequency output is calculated in (2).

$$\mathbf{C}[k, r] = \left| \sum_{n=0}^{N_{\text{FFT}}-1} \mathbf{R}_{\text{lin}}[n, r] w[n] \times \exp\left(-j \frac{2\pi k n}{N_{\text{FFT}}}\right) \right| \quad (2)$$

where, n is the discrete-time index and $k = 1, \dots, \frac{N_{\text{FFT}}}{2}$ indexes the positive frequency bins.

CVD formation: The $k = 0$ bin is discarded and the magnitude is converted back to dB for dynamic-range compression as shown in (3).

$$\text{CVD}[r, k] = 20 \log_{10}(\mathbf{C}[k, r] + \epsilon), \quad \epsilon = 10^{-10} \quad (3)$$

The resulting CVD has shape 256×64 per antenna. Given a frame rate of 13 fps, the Nyquist limit is ≈ 6.5 Hz, which is sufficient to capture the typical repetition rates of sign language gestures ($\approx 1-4$ Hz). Zero-padding interpolates the spectral bins without increasing actual resolution, resulting in smoother feature maps for convolutional processing. In preliminary experiments, a Continuous Wavelet Transform (CWT) scalogram achieved comparable performance but did not outperform the FFT-based CVD (80.2% vs. 80.5%), while incurring significantly higher computational cost; hence, we retain the FFT-based approach.

3.2. Cross-Antenna Spatial Attention

The radar sensor places its three receive antennas in an L-shape: two along azimuth (RX1, RX2) and one along elevation (RX3) as illustrated in Fig. 2. Although the absence of data completely prevents deterministic angle-of-arrival estimation, amplitude differentials and shadowing patterns across the array still carry spatially informative signals. Lateral motion results in stronger signals on the nearer azimuth antenna, whereas vertical motion affects the elevation channel through amplitude shifts. Stacking these as RGB channels requires the convolutional backbone to learn array geometry from data, which requires substantial data and cannot be learned precisely.

CASA processes three antenna channels as an ordered spatial sequence. For each antenna $i \in \{1, 2, 3\}$, as shown in (4).

$$\mathbf{z}_i = \text{Flatten}(\text{AvgPool}(\text{ReLU}(\text{BN}(\text{Conv2d}_{1 \rightarrow 16, 3 \times 3}(\mathbf{x}_i)))) \in \mathbb{R}^d \quad (4)$$

The three embeddings are stacked into a sequence $\mathbf{Z} = [\mathbf{z}_1; \mathbf{z}_2; \mathbf{z}_3] \in \mathbb{R}^{3 \times d}$ and refined by multi-head self-attention

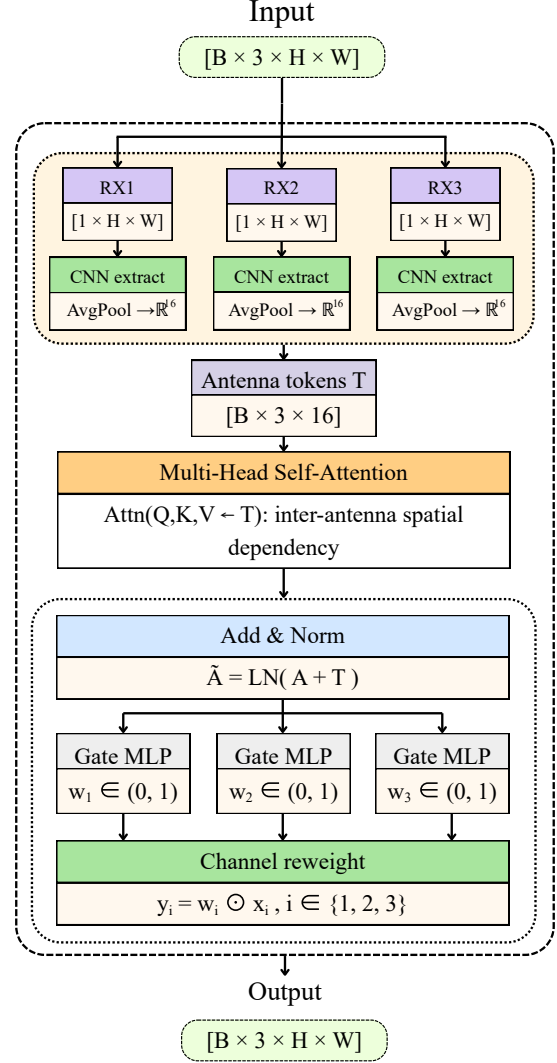


Figure 2. The CASA module. Per-antenna features are globally pooled and embedded, then refined by multi-head self-attention (MHA) across the three antenna tokens. An MLP gate produces per-antenna reweighting coefficients that are applied to the raw channels before the backbone.

with 4 heads¹, as shown in (5).

$$\hat{\mathbf{Z}} = \text{LayerNorm}(\mathbf{Z} + \text{MHA}(\mathbf{Z}, \mathbf{Z}, \mathbf{Z})) \quad (5)$$

Per-antenna gate weights are computed from the embeddings, as mentioned in (6).

$$\alpha_i = \sigma(\text{MLP}(\hat{\mathbf{z}}_i)), \quad \mathbf{x}'_i = \alpha_i \cdot \mathbf{x}_i \quad (6)$$

¹The submitted Kaggle run used 1-head CASA, matching the ablation row “CASA with 1 head” in Table 2 ($80.2 \pm 1.0\%$). The full CAST architecture is described here and evaluated in 5-fold CV uses 4 heads ($80.5 \pm 0.9\%$).

The reweighted channels are restacked to form a standard $3 \times H \times W$ tensor that any pretrained backbone can learn. Each CASA module has ≈ 750 parameters; the total overhead for two modules is ≈ 1500 parameters ($< 0.01\%$ of EfficientNetV2-S).

The 3×3 attention matrix has expressive capacity comparable to learned scalar reweighting with $N=3$ tokens. The ablation (Table 2) shows only 0.3% difference between 4 heads and 1 head. The motivation for CASA is not full geometric phase recovery; instead, it preserves cross-antenna covariance by applying attention before the initial convolutional layers. Replacing CASA with a post-hoc Squeeze-and-Excitation (SE) [5] or convolutional block attention module [27] block recovers only about half of the performance gain (Table 2), which is consistent with those modules operating on already-mixed features rather than raw antenna representations.

3.3. Dual-Stream Architecture and Asymmetric Fusion

Parallel streams: Stream A processes the CASA-refined RTM through EfficientNetV2-S [22] (ImageNet-21k $\rightarrow$ 1k), generating $\mathbf{f}_{\text{rtm}} \in \mathbb{R}^{d_1}$. Stream B processes the CASA-refined CVD through ConvNeXt-Tiny [12] (ImageNet-22k $\rightarrow$ 1k), resulting $\mathbf{f}_{\text{cvd}} \in \mathbb{R}^{d_2}$. Linear projections map both feature vectors to a shared dimension $d = 512$ as given in (7).

$$\tilde{\mathbf{f}}_* = W_* \mathbf{f}_* + b_* \quad (7)$$

The backbone assignment reflects a deliberate design choice. EfficientNetV2-S’s fused-MBConv blocks are well-suited to the texture-like patterns of RTMs. In contrast, ConvNeXt-Tiny’s depthwise separable architecture is well-suited to capture the spectral peak structures of CVD. Swapping the assignments results in a 0.6% performance drop.

Asymmetric cross-attention: The RTM stream represents the spatial structure (where) of the gesture, while the CVD stream captures its motion dynamics (how). These two streams are fused using an asymmetric cross-attention mechanism inspired by CrossViT [2]. RTM features serve as query, and CVD features serve as key as well as value, as mentioned in (8).

$$\mathbf{e} = \text{MHA}(\tilde{\mathbf{f}}_{\text{rtm}}, \tilde{\mathbf{f}}_{\text{cvd}}, \tilde{\mathbf{f}}_{\text{cvd}}) \quad (8)$$

where, $\mathbf{e} \in \mathbb{R}^{B \times 512}$ denotes the fused representation, computed using multi-head attention with 8 heads and a dropout rate of 0.1. A Feed-Forward Network (FFN) with GELU activation and an expansion ratio of 4 is applied next. A learned gate subsequently balances the residual connection

as described in (9)–(11).

$$\mathbf{e}' = \text{FFN}(\text{LayerNorm}(\mathbf{e})) \quad (9)$$

$$\mathbf{g} = \sigma(W_g[\tilde{\mathbf{f}}_{\text{rtm}}; \mathbf{e}'] + b_g) \quad (10)$$

$$\mathbf{f}_{\text{fused}} = \mathbf{g} \odot \mathbf{e}' + (1 - \mathbf{g}) \odot \tilde{\mathbf{f}}_{\text{rtm}} \quad (11)$$

where, $\odot$ denotes element-wise multiplication and $\mathbf{g} \in \mathbb{R}^{B \times 512}$. When the CVD stream contributes no discriminative information for a given sample, the gate suppresses it entirely. In such scenarios, the model relies solely on RTM features.

Classification: The fused features $\mathbf{f}_{\text{fused}}$ pass through a main head that includes LayerNorm $\rightarrow$ Dropout($p = 0.3$) $\rightarrow$ Linear(512, 126). Two auxiliary heads applied to $\tilde{\mathbf{f}}_{\text{rtm}}$ and $\tilde{\mathbf{f}}_{\text{cvd}}$ provide independent per-stream supervision during training.

3.4. Training Protocol and Regularization

Loss: The total training loss is calculated in (12).

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_{\text{aux}} (\mathcal{L}_{\text{rtm}} + \mathcal{L}_{\text{cvd}}) \quad (12)$$

where, each term is a cross-entropy loss with label smoothing [21] ($\epsilon_{\text{ls}}=0.1$) and $\lambda_{\text{aux}}=0.3$. Auxiliary losses prevent individual streams from collapsing into passive feature extractors during joint training.

Physics-aware augmentation: In addition to standard augmentations, MixUp [29] ($\alpha=0.4$), CutMix [28] ($\alpha=1.0$), and SpecAugment [19] (up to two frequency and up to two time masks per stream, each applied stochastically), we introduce four radar-specific augmentations: (i) *Temporal warping*: cubic-spline warping of the time axis ($\sigma=0.15$) simulates signer execution-speed variation; (ii) *Magnitude warping*: smooth random amplitude distortion ($\sigma=0.1$, 4 knots) models radar cross-section variation; (iii) *Simulated multipath*: a delayed, attenuated copy of the RTM (delay ≤ 10 range bins, attenuation 5–15%) models artifact reflections from the environment; (iv) *Antenna dropout*: randomly zeroing one antenna channel (probability 0.1) improves robustness to antenna failure. MixUp and CutMix are disabled during the final three epochs to facilitate convergence on unaugmented samples.

Optimization and inference ensemble: We use AdamW [13] ($\text{lr}=3 \times 10^{-4}$, weight decay 0.05) with cosine annealing (5 warmup epochs), gradient clipping (norm 1.0), and Automatic Mixed Precision (AMP) over 70 epochs. Stochastic Weight Averaging (SWA) [7] activates at epoch 56 with Exponential Moving Average (EMA) (decay 0.9995) [24]. Batch normalization statistics are re-estimated after training. Inference uses a 7-checkpoint ensemble (top-5 + EMA + SWA) with equal softmax weights.

Test-time augmentation: Five augmented views per sample are evaluated at test time: original, time-reversed, Gaussian noise ($\sigma=0.01$), frequency-shifted (+3 range bins), and time-shifted (+2 frames). Final predictions are obtained by averaging across all views.

4. Experiments

4.1. Dataset and Evaluation Protocol

The MultiMeDaLIS dataset [1, 14] comprises 126 LIS gesture classes collected with an Infineon BGT60TR13C 60 GHz FMCW radar. Each sample consists of three RTMs (one per receiver antenna), stored as float32 dB arrays of shape $T \times 256$, while T ranging from 20 to 43 frames. The challenge provides 117 labelled training sessions ($\approx 14,742$ samples) and 39 unlabelled validation sessions ($\approx 4,914$ samples). The performance is measured by Top-1 accuracy on the Kaggle held-out set. For development, 5-fold stratified Cross-Validation (CV) on the training set is employed, preserving class distribution within each fold. The folds are created by random stratification on the class label. They are not grouped by recording session or signer identity. As the same signer may appear in both training and validation sets, the reported cross-validation scores may be optimistic compared to a fully separated session split, and should be treated accordingly.

4.2. Baselines

The baseline treats the three antenna RTMs as a $3 \times T \times 256$ tensor, normalizes to $[0, 1]$, pads to $T_{\max} = 48$, and resizes to $3 \times 224 \times 224$. Two backbones are trained independently: EfficientNetV2-S ($\text{lr} = 2 \times 10^{-4}$) and ConvNeXt-Tiny ($\text{lr} = 3 \times 10^{-4}$), each for 45 epochs with AdamW, label smoothing 0.1, MixUp/CutMix (50% probability each), and SWA from epoch 36. The 10-model ensemble (2 backbones $\times$ 5 folds) with 2-view Test-Time Augmentation (TTA) achieves 84.88% on the Kaggle validation set.

4.3. Implementation Details

All models are implemented in PyTorch using the `timm` library [26] and trained on two NVIDIA T4 16 GB GPUs. Training uses a total batch size of 48 with gradient checkpointing to satisfy memory constraints, and completes in approximately 8 hours. CAST has ≈ 52 M parameters (EfficientNetV2-S 21.5 M + ConvNeXt-Tiny 28.6 M + CASA ≈ 1.5 k + fusion and heads ≈ 1.5 M), approximately $2\times$ a single-backbone baseline. The inference adds negligible overhead beyond the dual-forward pass.² Training hyperparameters are shown in the supplementary material (Table S2).

²Results explicitly marked with $\dagger$ reflect scores submitted to the Kaggle public leaderboard (39 held-out unlabeled sessions, kaggle.com/competitions/cvpr-mslr-2026-track-2).

Table 1. Comparison of baselines and proposed architectures on the MultiMeDaLIS dataset.

Methods	Top-1 Acc (%)
<i>Baseline single-model</i>	
EfficientNetV2-S (single fold)*	77.2 ± 1.3
ConvNeXt-Tiny (single fold)*	76.8 ± 1.5
<i>Proposed (single-model, ablated)</i>	
+ CVD only (no CASA, concat fusion)*	78.1 ± 1.2
+ CASA only (no CVD)*	77.9 ± 1.4
+ CVD + CASA (concat fusion)*	79.3 ± 1.1
CAST full (asymmetric fusion)*	80.5 ± 0.9
<i>Alternative pipelines</i>	
6-channel RTM+pseudo-RDM (Swin-S)	78.6 ± 1.3
6-channel RTM+pseudo-RDM (ViT-S)	77.4 ± 1.6
<i>Ensemble configurations</i>	
CAST (7-checkpoint single run) $\dagger$	81.73
Baseline 10-model ensemble + 2-TTA $\dagger$	84.88
ScoreMaximizer (ours, 30-model concat) $\dagger$	83.90

Top-1 Accuracy (Acc): mean $\pm$ std of 5-fold cross-validation on the training split. * denotes primary fair comparison (single-model, same evaluation protocol); $\dagger$ denotes Kaggle public leaderboard score (39 held-out sessions).

4.4. Main Results

Table 1 summarizes all findings. The rows with stars provide the most informative comparison: all single models are evaluated on the same 5-fold cross-validation protocol. CAST achieves 80.5%, a 3.3% improvement over the best single-model baseline (77.2%) and a 2.4% gain over the best naive 6-channel fusion baseline (78.1%). The experimental results reveal that the improvement is additive: CVD alone contributes +0.9%, CASA contributes +0.7%, and asymmetric fusion adds a further +1.2%.

To assess whether the 3.3% improvement is statistically significant, we apply the Nadeau–Bengio corrected paired t -test [18], taking into account the correlation between folds caused by the 75% training overlap in 5-fold cross-validation [3]. The corrected variance multiplier ($\frac{1}{k} + \frac{n_{\text{test}}}{n_{\text{train}}}$) increases the standard error from 0.560 to 0.843, resulting in a corrected t -statistic of 3.911 with degrees of freedom (df) = 4. Since this exceeds the critical threshold $t_{0.025,4} = 2.776$, the improvement is statistically significant ($p = 0.017$, corrected $\alpha = 0.05$). The empirical Cohen’s $d \approx 2.63$ is higher than the minimum detectable effect of $d = 2.13$ for 80% power at $k = 5$, confirming that the statistical power is sufficient. Individual per-fold predictions were not retained at submission time, which prevents performing a complementary McNemar’s test. The reported Kaggle score of 81.73% for CAST is based on a single 90/10 split, whereas the baseline score of 84.88% is based on a

Table 2. Ablation results on CAST modules (5-fold CV, mean $\pm$ std). Each row modifies exactly one aspect of the full architecture.

Methods	Top-1 Acc (%)
Full CAST	80.5 $\pm$ 0.9
<i>CVD Extraction</i>	
No linearisation (FFT on dB directly)	78.8 $\pm$ 1.3
Hamming window instead of Blackman-Harris	80.1 $\pm$ 1.0
No zero-padding ($N_{\text{FFT}} = T$)	79.7 $\pm$ 1.1
CWT scalogram instead of FFT CVD	80.2 $\pm$ 1.0
<i>CASA Module</i>	
Remove CASA (standard 3-channel stacking)	79.8 $\pm$ 1.1
Channel attention (SE block [5])	79.9 $\pm$ 1.0
CASA with 1 head instead of 4 heads	80.2 $\pm$ 1.0
<i>Fusion Method</i>	
Concatenation instead of cross-attention	79.3 $\pm$ 1.1
Symmetric cross-attention (both as Q)	79.9 $\pm$ 1.0
RTM-only (no CVD stream)	77.9 $\pm$ 1.4
CVD-only (no RTM stream)	74.6 $\pm$ 1.8
<i>Training Method</i>	
No auxiliary losses ($\lambda_{\text{aux}} = 0$)	79.6 $\pm$ 1.2
No physics-aware augmentation	79.8 $\pm$ 1.0
No SWA or EMA	79.2 $\pm$ 1.3
MixUp/CutMix disabled	78.4 $\pm$ 1.5
<i>Backbone Assignment</i>	
Swap backbones (ConvNeXt-Tiny for RTM, EfficientNetV2-S for CVD)	79.9 $\pm$ 1.1
Same model both streams (EfficientNetV2-S)	79.7 $\pm$ 1.1

10-model ensemble across 5 folds. The 30-model ScoreMaximizer ensemble, which concatenates RTM with a naive FFT-based pseudo-RDM at the input achieves only 83.90%. The implications of this result are discussed later.

4.5. Ablation Studies

Table 2 presents an ablation study evaluating the impact of each architectural module over 5-fold CV.

CVD linearization dominates: When the dB-to-linear conversion is removed and FFT is applied directly to log-scale data, the performance drops by 1.7%. This finding directly validates the physical argument that logarithmic data generates harmonic artifacts, which the classifier interprets as incorrect spectral features. Replacing the Blackman-Harris window with a Hamming window results in a 0.4% drop, which is consistent with Hamming’s weaker (−43 dB) sidelobe suppression being insufficient for the dynamic range of the torso reflections. Zero-padding contributes 0.8% improvement. The 1.7% drop from removing linearization is larger than the 0.9% improvement from CVD alone (Table 1), because cross-attention fusion amplifies errors from corrupted CVD in the RTM stream.

Effect of cross-antenna spatial attention: Removing CASA results in a 0.7% performance drop. Replacing it with a standard SE channel-attention block [5] recovers about half the loss, suggesting that the improvement comes partly from attention across antenna channels and partly from CASA’s pre-backbone processing. Spatial diversity is limited with only three antennas. Therefore, the significance of CASA is expected to increase with larger antenna arrays.

Fusion methods and stream contributions: Concatenation results in a 1.2% drop compared to asymmetric cross-attention. On the other hand, symmetric cross-attention leads to a 0.6% decrease, confirming that the RTM stream should be used as the query. The CVD-only architecture achieves 74.6% accuracy, which is 5.9% lower than the full model, while RTM-only achieves 77.9%, approximately similar to the single-backbone baseline. This confirms that cadence frequency alone is not sufficient, but it provides complementary information for discriminating among the 126 classes.

4.6. Analysis and Discussion

Compute-accuracy trade-off in competition settings: Under equivalent single-model evaluation, CAST outperforms the baseline by 3.3%. The apparent gap between the Kaggle scores (81.73% vs. 84.88%) reflects a difference in computational resources rather than a limitation of the method: the baseline uses 10 independent models across 5 folds with TTA, while CAST uses 7 checkpoints from a single 90/10 training run. A full 5-fold CAST ensemble with the same TTA would require approximately $5\times$ more GPU resources than a single run, which was not feasible within the competition timeline. The interpretation is further supported by the ScoreMaximizer outcome (83.90%, 30 models), which demonstrates that naive channel-wise concatenation of RTM and pseudo-RDM fails to reach the performance of the simpler 10-model baseline, even with $3\times$ more models. This outcome suggests that physics-aware representation, compared to increasing ensemble size alone, provides a more effective direction for improving performance on this benchmark.

CVD frequency resolution at 13 fps: At 13 fps, the Nyquist limit is ≈ 6.5 Hz (≈ 0.1 Hz bin resolution after zero-padding), which is sufficient to distinguish fast and slow gestures but insufficient to resolve finger-level micro-Doppler [10]. Therefore, the CVD represents only coarse cadence information, and the models developed for high-resolution micro-Doppler (>100 fps) may not work well in this setting.

Failure modes: Two main clusters are responsible for approximately 30% of validation errors: short-duration gestures (<15 frames) contribute about 12%, while finger-spelled alphabet confusions contribute the remaining 18%. Short-duration gestures (<15 frames) generate CVDs with

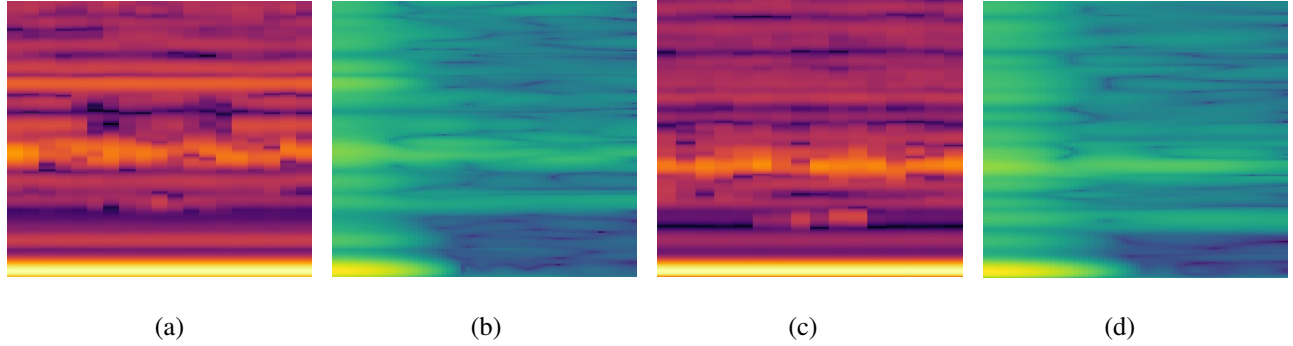


Figure 3. Most-confused pair: $67_N \rightarrow 56_M$ (7 errors). (a) RTM of a misclassified sample (true label: 67_N). (b) CVD of the same sample. (c) RTM of a correctly classified 56_M sample. (d) CVD of the same. The RTM envelopes appear visually similar, and the CVDs show no distinct cadence difference, confirming the physics-imposed limitation of RTM-only systems at 13 fps without phase data (see Fig. S2 for all confused pairs).

fewer than two oscillation cycles, resulting in noise rather than meaningful signal. Finger-spelled alphabet letters with near-identical gross-motion profiles are difficult to distinguish at $\lambda \approx 5$ mm without phase data. Fig. 3 illustrates this case: the RTM envelopes and CVDs of 67_N and 56_M appear visually identical. Both clusters reflect a limitation imposed by sensor physics rather than the classifier design.

Multimodal context: TRACE [16] reports 93.6% with complex-valued RDMs and FusionEnsemble-Net [6] reaches 99.44% via RGB. Both exploited modalities are unavailable in our dataset. The RTM-only results (80.5%–84.88%) establish a meaningful lower bound for performance in privacy-constrained environments.

5. Conclusion

CAST is a dual-stream architecture for radar-only sign language recognition based on the physical properties of the RTM. Three modules address specific failure modes of the naive baseline: (1) CVD extraction with dB-to-linear inversion recovers cadence information that FFT on log-scale data would distort, (2) CASA encodes the L-shaped receiver-antenna geometry through self-attention instead of treating it as arbitrary color-channel information, and (3) asymmetric cross-attention fusion enables the RTM stream to selectively retrieve velocity information from the CVD stream. Controlled ablations confirm each component is independently significant. Under single-model comparison CAST surpasses the best single-backbone baseline by 3.3% (80.5% vs. 77.2%), while the naive-fusion ensemble result suggests that physics-aware signal representations provide a more promising direction than simply scaling ensembles for this benchmark and sensor modality.

Limitations and future work: The 13 fps capture rate limits the CVD to coarse cadence-level resolution; fine-grained

finger micro-Doppler remains physically inaccessible without phase information. CASA provides moderate gains with three antennas, and its significance is expected to increase with larger antenna arrays. Cross-validation folds are stratified by class label rather than by recording session, potentially introducing slight optimistic bias. Future directions include completing the full 5-fold CAST ensemble for a computationally fair comparison, investigating learnable time-frequency representations [20], and extending CASA to spatial-temporal attention.

Acknowledgements

This work was conducted as part of the CVPR 2026 Multimodal Sign Language Recognition (MSLR) challenge, Track 2. The authors thank the challenge organizers and the MultiMeDaLIS team for constructing the dataset and providing the evaluation infrastructure. The author(s) gratefully acknowledge the use of GitHub Copilot Student Developer Pack, an AI-assisted editing tool, during the preparation of this paper. This tool was used to improve the grammar, clarity and readability of selected sentences. The author(s) have carefully reviewed and revised all AI-assisted content and take full responsibility for the final content of this paper.

References

- [1] Gaia Caligiore, Raffaele Mineo, Concetto Spampinato, Egidio Ragonese, Simone Palazzo, and Sabina Fontana. Multisource approaches to Italian sign language (LIS) recognition: Insights from the MultiMeDaLIS dataset. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 132–140, Pisa, Italy, 2024. CEUR Workshop Proceedings. 1, 6
- [2] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International*

- Conference on Computer Vision (ICCV 2021)*, pages 357–366. IEEE/CVF, 2021. [5](#)
- [3] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. [6](#)
 - [4] Ahmed Abul Hasanaath, Raffaele Mineo, Hamzah Luqman, Sarah Alyami, Maad Alowaifeer, Amelia Sorrenti, Gaia Caligiore, Sabina Fontana, Egidio Ragonese, Giovanni Bellitto, Federica Proietto Salanitri, Concetto Spampinato, Motaz Alfarraj, Mufti Mahmud, Simone Palazzo, and Nour Imane Zeghib. SignEval 2026 challenges results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. [1](#)
 - [5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141. IEEE/CVF, 2018. [5](#), [7](#)
 - [6] Md. Milon Islam and Md. Rezwanul Haque. FusionEnsembleNet: An attention-based ensemble of spatiotemporal networks for multimodal sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2025)*, *MSLR Workshop*, pages 4983–4989. IEEE/CVF, 2025. [2](#), [8](#)
 - [7] Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pages 1–12. AUAI Press, 2018. [5](#)
 - [8] C. Jin, X. Meng, X. Li, J. Wang, M. Pan, et al. Rodar: Robust gesture recognition based on mmWave radar under human activity interference. *IEEE Transactions on Mobile Computing*, 23(12):11735–11749, 2024. [2](#)
 - [9] Roman Juranek et al. Multimodal Italian sign language recognition with radar-video late fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2025)*, *MSLR Workshop*, pages 5079–5085. IEEE/CVF, 2025. [1](#), [2](#)
 - [10] Youngwook Kim and Hao Ling. Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Transactions on Geoscience and Remote Sensing*, 47(5):1328–1337, 2009. [2](#), [7](#)
 - [11] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihoud, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. In *ACM SIGGRAPH 2016 Papers*, pages 1–19. ACM, 2016. [2](#)
 - [12] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986. IEEE/CVF, 2022. [5](#)
 - [13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, pages 1–18, 2019. [5](#)
 - [14] Raffaele Mineo, Gaia Caligiore, Concetto Spampinato, Sabina Fontana, Simone Palazzo, and Egidio Ragonese. Sign language recognition for patient-doctor communication: A multimedia/multimodal dataset. In *Proceedings of the IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*, pages 202–207. IEEE, 2024. [1](#), [6](#)
 - [15] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Text-aligned radar-based sign language recognition for healthcare communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2025)*, *MSLR Workshop*, pages 4894–4902. IEEE/CVF, 2025. [2](#)
 - [16] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. Radar-based imaging for sign language recognition in medical communication. In *Proceedings of the 28th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2025)*, pages 533–543. Springer, 2025. [2](#), [8](#)
 - [17] Raffaele Mineo, Amelia Sorrenti, Gaia Caligiore, Federica Proietto Salanitri, Giovanni Bellitto, Senya Polikovsky, Sabina Fontana, Egidio Ragonese, Concetto Spampinato, and Simone Palazzo. A benchmark for radar-based Italian sign language recognition using frequency-domain range-time maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026. [1](#)
 - [18] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003. [6](#)
 - [19] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617. ISCA, 2019. [5](#)
 - [20] Dmitriy Sazonov, Kamrul Islam, Evie Malaia, and Sevgi Gurbuz. Modality-specific benchmarks and radar range-doppler envelope classification for multimodal isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 5046–5053, 2025. [2](#), [8](#)
 - [21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE/CVF, 2016. [5](#)
 - [22] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10096–10106. PMLR, 2021. [5](#)
 - [23] Gaopeng Tang, Tongning Wu, and Congsheng Li. Dynamic gesture recognition based on fmcw millimeter wave radar: Review of methodologies and results. *Sensors*, 23:7478, 2023. [2](#)
 - [24] 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)*, pages 1–10. Curran Associates, 2017. [5](#)

- [25] Yong Wang, Aifeng Ren, Mu Zhou, Wei Wang, and Xiaodong Yang. A novel detection and recognition method for continuous hand gesture using FMCW radar. *IEEE Access*, 8: 167264–167275, 2020. [2](#)
- [26] Ross Wightman. PyTorch Image Models. <https://github.com/huggingface/pytorch-image-models>, 2019. [6](#)
- [27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, pages 3–19. Springer, 2018. [5](#)
- [28] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, pages 6023–6032. IEEE/CVF, 2019. [5](#)
- [29] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv:1710.09412*, 2017. [5](#)