Transfer Learning from Text-Motion Retrieval to Violence Classification

Leveraging Semantic Motion Representations for Action Recognition

Alessio Pannozzo
pannozzo.1960374@studenti.uniroma1.it

Federico Raponi
raponi.1963339@studenti.uniroma1.it

Marco Realacci
marco.realacci@uniroma1.it

Lorenzo Spataro
spataro.1946590@studenti.uniroma1.it

Abstract

Violence detection from skeletal motion is critical for safety applications but challenging due to extreme class imbalance. We explore whether Text-Motion Retrieval (TMR++) representations transfer to violence classification on BABEL-120, using prototypical networks with MAML adaptation. In 30-shot evaluation, we achieve 58.5% mAP with 73.1% Violence Detection Rate at 1.0% False Positive Rate, while fully supervised fine-tuning reaches 68.3% mAP.

1 Task and Motivation

Detecting violent actions from 3D skeletal motion is essential for surveillance, content moderation, and assistive robotics [1]. However, violence detection faces a fundamental challenge: violent actions are rare events. In typical action datasets, violent classes contain only 10-20 examples while common actions have thousands, creating severe class imbalance that causes standard supervised methods to ignore minority classes.

This drives the development of few-shot learning approaches, which enable models to adapt to rare classes with just a few labeled examples [2, 3].

2 Related Work

2.1 Cross-Modal Motion Representation

Recent advances in cross-modal learning leverage Transformer-based architectures to align human motion with natural language. Among these approaches, **TMR++** [4], an improvement over TMR [5], establishes *state-of-the-art* performance in text-to-motion retrieval through a dual-encoder framework trained with contrastive learning. Its architecture evolves from **TEMOS** [6] and **ACTOR** [7], adapting their Transformer VAE backbones from generative tasks to retrieval by optimizing a shared motion-text latent space.

Our hypothesis is that this text-aligned space provides a robust semantic prior for detecting rare safety-critical actions.

2.2 Few-Shot Learning and MAML

Few-shot learning is essential for violence detection given the scarcity of labeled data. Metric-based approaches like Prototypical Networks [3] perform classification by measuring distances to class prototypes in a learned embedding space. Conversely, optimization-based methods like Model-Agnostic Meta-Learning (MAML) [2] learn an optimal parameter initialization that adapts to new tasks via minimal gradient updates. Our work combines these paradigms by applying MAML adaptation to the semantic motion embeddings from TMR++.

3 Methodology

3.1 Model Architecture

Building on this text-aligned space, our approach leverages semantic motion representations to detect violence in skeletal data. We utilize the **TMR++** motion encoder as our backbone. TMR++ is a dual-encoder architecture originally designed for text-motion retrieval. It employs an ACTOR-style [7] Transformer encoder consisting of 6 Transformer layers with 4 attention heads.

The encoder f_θ takes normalized 3D motion sequences (263-dimensional joint features at 20 Hz) and maps them to a d-dimensional latent embedding space ($d = 256$).

Depending on the training paradigm, we utilize two different classification heads:

- **Meta-Learning Head:** We use the encoder outputs directly to build prototypes for non-parametric classification during the MAML adaptation process.

- **Supervised Projection Head:** For our fully supervised baseline, we append a Multi-Layer Perceptron (MLP) to the encoder. This MLP projects the 256-dimensional input into a 1024-dimensional hidden space, compresses it to 512 dimensions, and maps it to class scores. Each linear transformation utilizes Layer Normalization, GELU activation, and Dropout for regularization.

3.2 Meta-Learning Framework (MAML)

We formulate violence detection as a few-shot multi-label classification problem. To enable rapid adaptation to rare violent classes with minimal labeled examples, we employ Model-Agnostic Meta-Learning (MAML) [2].
MAML learns an initialization of model parameters θ that is optimally positioned for fast adaptation to new tasks through fine-tuning. The meta-learning procedure operates through a two-stage optimization:

3.2.1 Inner Loop (Task Adaptation)

For a sampled task $\mathcal{T}$ (a set of action classes), the model creates a support set $\mathcal{D}^{sup}$. We perform $T = 20$ gradient steps on this support set to generate adapted parameters θ' specific to that task:

$$\theta' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}}^{sup}(\theta) \tag{1}$$

where $\alpha = 0.02$ is the inner loop learning rate.

3.2.2 Outer Loop (Meta-Optimization)

After adaptation, the model is evaluated on a held-out query set $\mathcal{D}^{qry}$ from the same task. We then compute the gradient of this query loss with respect to the *original* parameters θ. Due to resource constraints, we only utilize First-Order MAML (FOMAML) [8] and ignore second-order derivatives.

3.3 Optimization Objectives

To ensure the model adapts to violent classes without losing general action recognition capabilities, we introduce a combined **Three-Loss Architecture** during the inner-loop adaptation:

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \lambda_1 \mathcal{L}_{class} + \lambda_2 \mathcal{L}_{violence} \tag{2}$$

- **BCE Loss ($\mathcal{L}_{BCE}$):** A standard Binary Cross-Entropy loss for multi-label classification.
- **Center Loss ($\mathcal{L}_{class}$):** Penalizes the distance between samples and their class prototypes, as defined in [9].

$$\mathcal{L}_{class} = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|S_c|} \sum_{i \in S_c} (1 - s(\mathbf{z}_i, \mathbf{p}_c)) \tag{3}$$

- **Hinge Loss ($\mathcal{L}_{violence}$):** A margin-based loss that explicitly enforces separation between safety-critical categories. It maximizes the cosine distance between the centroid of all violent prototypes ($\mathbf{c}_V$) and the centroid of non-violent prototypes ($\mathbf{c}_N$), as defined in [10].

$$\mathbf{c}_V = \frac{1}{|V|} \sum_{i \in V} \mathbf{p}_i, \quad \mathbf{c}_N = \frac{1}{|N|} \sum_{j \in N} \mathbf{p}_j \tag{4}$$

$$\mathcal{L}_{violence} = \max(0, 1 - d(\mathbf{c}_V, \mathbf{c}_N)) \tag{5}$$

We empirically determine the optimal weighting coefficients through grid search, obtaining $\lambda_1 = 0.5$ and $\lambda_2 = 0.7$.

3.4 Training Strategies

3.4.1 Violence-Aware Episodic Sampling

Standard uniform sampling in meta-learning is insufficient for our task because violent classes are extremely rare ($< 5\%$ of classes). To ensure the model effectively learns to identify violence, we implement a biased sampling strategy:

1. We partition the dataset into "Violent" and "Non-Violent" sets defining 6 violent action categories (*kick, punch, hit, martial art, evade, grab person*).

2. In every training episode, we force the inclusion of violent classes by sampling $\rho = 40\%$ of the N classes from the Violent set.

This ensures that every meta-update includes gradients related to safety-critical concepts.

3.4.2 Fully Supervised Baseline

To establish a performance ceiling, we also implement a traditional fully supervised training regime. We fine-tune the TMR++ encoder and the attached MLP head on the complete BABEL-120 dataset using the same Three-Loss architecture described in Section 3.3.

4 Dataset and Benchmark

4.1 HumanML3D, Kit-ML

The TMR encoders utilized in our architecture are pretrained on large-scale motion-language datasets. HumanML3D [11] and Kit-ML [12] contain 3D human motions paired with textual descriptions, used to learn fine-grained alignment between action semantics and motion sequences.

4.2 BABEL

BABEL [13] consists of *5,847* training and *1,911* test sequences, represented by *263-dimensional* SMPL joint features at 20 Hz, each sample is multi-labeled. The dataset exhibits severe class imbalance, with common actions like *walk* having 1,695 examples while rare violent actions. We utilize BABEL-120, a subset containing 120 semantic action labels.

HumanML3D and BABEL are built on AMASS, a unified SMPL representation of diverse motion capture sources.

4.3 Evaluation Protocols

Episodic N-Way K-Shot Few-Shot. This is the standard meta-learning evaluation benchmark [3, 2]. In each test episode, the model is presented with $N \in \{5, 7, 10\}$ random classes. It is given $K \in \{1, 3, 5\}$ labeled support examples per class to adapt, and is then evaluated on unseen query examples. Due to GPU memory constraints (24GB), we weren't able to test on bigger N and K.

Full 120-Class Evaluation. We also evaluate global MAML adaptation on all 120 classes simultaneously, using K-shot support samples per class from training data and evaluating on the complete test set (1,911 samples), processed in mini-batches of size 32 due to memory constraints.

Metrics: Mean Average Precision (mAP), Top-K Accuracy, Recall@K. For violence-specific evaluation, we report Violence Detection Rate (VDR) and False Positive Rate (FPR), computed at the threshold maximizing the F1 score.

5 Experimental Results

5.1 Episodic Few-Shot Performance

We first evaluate our approach on the standard N-way K-shot episodic benchmark. Table 1 presents the results comparing the standard Prototypical Network (TMR) against our MAML adaptation.

MAML vs. Non-Parametric. In terms of general classification (mAP), MAML provides consistent gains. For instance, in the 5-way 5-shot scenario, MAML improves Top-1 accuracy from 75.8% to 78.6% (+2.8%). While gains in lower-shot settings (e.g., 1-shot) are more modest, the trend shows that MAML effectively leverages support data to fine-tune the feature space.

Consistent Improvements in Violence Metrics. MAML adaptation frequently improves *both* the Violence Detection Rate (VDR) and the False Positive Rate (FPR). In the **5-way 5-shot** setting, MAML increases VDR from 85.4% to 88.1% while simultaneously reducing FPR significantly from 12.8% to 8.0%. This suggests that the gradient-based adaptation refines the decision boundary to be more discriminative, rather than simply shifting the bias towards the violent class.

5.2 Global 120-Class Evaluation

Table 2 presents performance on the full 120-class vocabulary, representing a realistic deployment scenario.

Scaling with Support Size. MAML demonstrates clear scalability. As the support set size increases to 30-shot, MAML achieves a robust **58.5% mAP**, outperforming the prototypical baseline of 51.6%.

High Sensitivity with Low False Alarms. The global evaluation highlights the model's safety capabilities.

- At **30-shot**, MAML achieves a high VDR of **73.1%** (compared to 67.7% for the baseline) while maintaining a negligible FPR of **1.0%**.
- In the **1-shot** setting, MAML improves VDR by over 4% (62.4% $\rightarrow$ 66.7%) while *halving* the FPR (6.0% $\rightarrow$ 3.0%).

This confirms that the meta-learned initialization generalizes well to the full action vocabulary, correctly identifying violent events without flooding the system with false alarms.

5.3 Ablation Studies

A critical challenge in training contrastive objectives for imbalanced data is the risk of **class collapse**, a well-documented phenomenon in supervised contrastive learning [14, 15].

The Collapse Phenomenon. Early iterations utilizing a standard combination of **BCE and a naive contrastive loss** (pulling all violent samples together) resulted in a degradation of fine-grained classification accuracy. This aligns with findings by Xue et al. [14], who showed that supervised contrastive learning is prone to collapsing representations of subclasses within a class by not capturing all their features. Although BCE aims to distinguish classes, the addition of a strong contrastive term that forced all "violent" actions into a single compact cluster overpowered the fine-grained semantics. This caused the model to merge distinct classes (e.g., "punch," "kick") into a generic "violence" concept, reducing Top-1 accuracy—a manifestation of what Xue et al. term "class collapse."

Our Three-Loss Architecture addresses this by decoupling the objectives, following principles from hierarchical contrastive learning [15]: $\mathcal{L}_{class}$ (Center Loss) [9] maintains intra-class compactness (preserving specific action semantics), while $\mathcal{L}_{violence}$ (Hinge Loss) acts solely on group centroids to enforce violence/non-violence separability without collapsing fine-grained distinctions. As shown in the supervised baselines (Table 2 bottom), the model maintains high Top-1 accuracy, successfully avoiding feature collapse while improving violence separability. This approach is consistent with recent work showing that

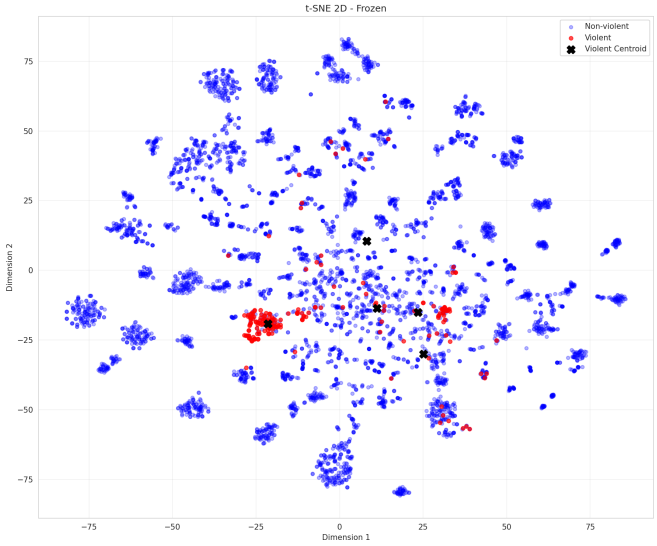

Figure 1: 2D t-SNE visualization of the latent space learned by the TMR model, k=5.

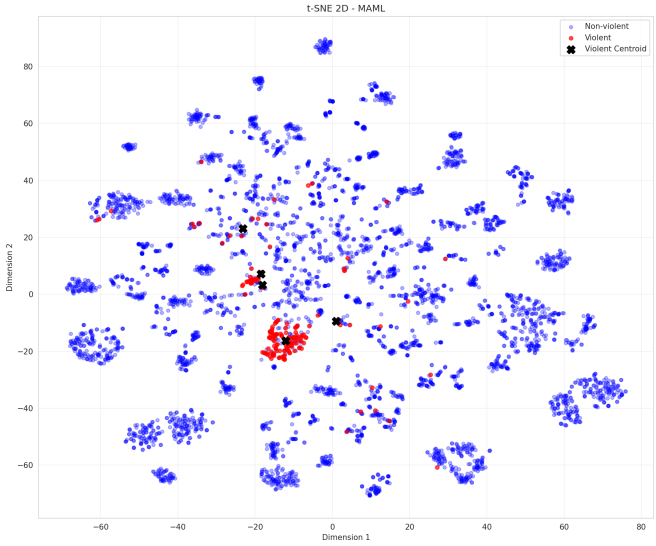

Figure 2: 2D t-SNE visualization of the latent space after MAML adaptation, k=5.

decoupled multi-level objectives can prevent class collapse in contrastive learning [14].

5.4 Qualitative Analysis

We visualize the impact of our adaptation strategy using t-SNE projections of the motion embeddings. Figures 1 and 2 show the embeddings before and after MAML adaptation, respectively, for the $k = 5$ scenario.

Compared to the baseline, the adapted model exhibits more compact within-class groupings and reduced overlap between classes. In particular, samples belonging to the same class tend to form tighter clusters, while different classes become more clearly separated after adaptation.

6 Conclusions and Limitations

In this work, we investigated the transferability of text-motion retrieval representations (TMR++) to the task of few-shot violence detection. By integrating Model-Agnostic Meta-Learning (MAML) with a violence-aware sampling strategy and a decoupled three-loss objective, we demonstrated that semantic priors can be effectively adapted to identify safety-critical actions.

Our experimental results lead to two primary conclusions. First, **MAML adaptation yields consistent improvements** over static prototypical networks. The gradient-based adaptation significantly enhances the model's ability to distinguish violent actions, achieving higher Violence Detection Rates (VDR) while simultaneously lowering False Positive Rates (FPR). This confirms that learning an adaptable initialization is superior to fixed metric-based approaches for this domain.

6.1 Limitations

Despite these gains, **a performance gap remains compared to fully supervised learning.** While our best 30-shot MAML configuration achieves a respectable 58.5% mAP, it still trails the fully supervised baseline (68.3% mAP) by approximately 10%. This indicates that while meta-learning is highly effective for data-scarce scenarios or rapid deployment to new violence types, it cannot yet fully replicate the robustness of a model trained on the complete data distribution.

6.2 Future Work

Future research could explore **semi-supervised meta-learning** [16], utilizing the abundant unlabeled motion data available in AMASS to regularize the few-shot adaptation process.

References

[1] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. "Violent Flows: Real-Time Detection of Violent Crowd Behavior". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2012.

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *International Conference on Machine Learning (ICML)*. 2017.

[3] Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical Networks for Few-shot Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[4] Léore Bensabath, Mathis Petrovich, and Gül Varol. "A Cross-Dataset Study for Text-based 3D Human Motion Retrieval". In: *European Conference on Computer Vision (ECCV)*. 2024.

[5] Mathis Petrovich, Michael J. Black, and Gül Varol. "TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis". In: *International Conference on Computer Vision (ICCV)*. 2023.

[6] Mathis Petrovich, Michael J. Black, and Gül Varol. "TEMOS: Generating diverse human motions from textual descriptions". In: *European Conference on Computer Vision (ECCV)*. 2022.

[7] Mathis Petrovich, Michael J. Black, and Gül Varol. "Action-Conditioned 3D Human Motion Synthesis with Transformer VAE". In: *International Conference on Computer Vision (ICCV)*. 2021.

[8] Alex Nichol, Joshua Achiam, and John Schulman. "On First-Order Meta-Learning Algorithms". In: *arXiv preprint arXiv:1803.02999* (2018).

[9] Yandong Wen et al. "A Discriminative Feature Learning Approach for Deep Face Recognition". In: *European Conference on Computer Vision (ECCV)*. 2016.

[10] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297.

[11] Chuan Guo et al. "Generating Diverse and Natural 3D Human Motions from Text". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.

[12] Matthias Plappert, Christian Mandery, and Tamim Asfour. "The KIT Motion-Language Dataset". In: *Big Data* 4.4 (2016), pp. 236–252.

[13] Abhinanda R. Punnakkal et al. "BABEL: Bodies, Action and Behavior with English Labels". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[14] Yihao Xue et al. "Which Features are Learnt by Contrastive Learning? On the Role of Simplicity Bias in Class Collapse and Feature Suppression". In: *International Conference on Machine Learning (ICML)*. 2023.

[15] Varsha Suresh and Desmond C. Ong. "Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2021.

[16] Mengye Ren et al. "Meta-Learning for Semi-Supervised Few-Shot Classification". In: *International Conference on Learning Representations (ICLR)*. 2018.

7 Appendix

Table 1: Few-Shot Performance with Violence Detection Metrics (mean ± std, 200 episodes)

Config	Model	mAP	Top-1	Top-5	Recall@1	Recall@5	VDR	FPR
5-w 0-s	TMR	$0.493^{\pm.014}$	$0.511^{\pm.018}$	$0.796^{\pm.012}$	$0.321^{\pm.016}$	$0.582^{\pm.014}$	$0.843^{\pm.016}$	$0.326^{\pm.033}$
7-w 0-s	TMR	$0.443^{\pm.009}$	$0.437^{\pm.011}$	$0.733^{\pm.011}$	$0.266^{\pm.009}$	$0.533^{\pm.010}$	$0.814^{\pm.016}$	$0.310^{\pm.032}$
10-w 0-s	TMR	$0.427^{\pm.008}$	$0.427^{\pm.009}$	$0.709^{\pm.010}$	$0.266^{\pm.008}$	$0.517^{\pm.010}$	$0.834^{\pm.012}$	$0.409^{\pm.029}$
5-w 1-s	TMR	$0.622^{\pm.009}$	$0.646^{\pm.014}$	$0.846^{\pm.008}$	$0.403^{\pm.010}$	$0.676^{\pm.008}$	$0.799^{\pm.012}$	$0.279^{\pm.028}$
	MAML	$\mathbf{0.624}^{\pm.010}$	$\mathbf{0.655}^{\pm.015}$	$\mathbf{0.860}^{\pm.009}$	$\mathbf{0.411}^{\pm.010}$	$\mathbf{0.681}^{\pm.009}$	$\mathbf{0.822}^{\pm.013}$	$\mathbf{0.200}^{\pm.019}$
7-w 1-s	TMR	$0.602^{\pm.010}$	$0.592^{\pm.013}$	$0.808^{\pm.011}$	$0.404^{\pm.012}$	$\mathbf{0.668}^{\pm.011}$	$\mathbf{0.790}^{\pm.017}$	$0.167^{\pm.021}$
	MAML	$\mathbf{0.608}^{\pm.011}$	$\mathbf{0.610}^{\pm.013}$	$\mathbf{0.811}^{\pm.010}$	$\mathbf{0.416}^{\pm.012}$	$0.666^{\pm.010}$	$0.774^{\pm.015}$	$\mathbf{0.093}^{\pm.012}$
10-w 1-s	TMR	$0.606^{\pm.008}$	$0.587^{\pm.010}$	$0.772^{\pm.009}$	$0.396^{\pm.009}$	$0.655^{\pm.011}$	$0.775^{\pm.011}$	$0.156^{\pm.018}$
	MAML	$\mathbf{0.607}^{\pm.008}$	$\mathbf{0.598}^{\pm.010}$	$\mathbf{0.798}^{\pm.007}$	$\mathbf{0.397}^{\pm.008}$	$\mathbf{0.669}^{\pm.009}$	$\mathbf{0.787}^{\pm.010}$	$\mathbf{0.139}^{\pm0.14}$
5-w 3-s	TMR	$0.752^{\pm.008}$	$0.743^{\pm.010}$	$0.902^{\pm.006}$	$0.488^{\pm.012}$	$0.794^{\pm.008}$	$0.826^{\pm.014}$	$\mathbf{0.062}^{\pm.014}$
	MAML	$\mathbf{0.764}^{\pm.009}$	$\mathbf{0.778}^{\pm.009}$	$\mathbf{0.915}^{\pm.007}$	$\mathbf{0.512}^{\pm.012}$	$\mathbf{0.800}^{\pm.009}$	$\mathbf{0.852}^{\pm.012}$	$0.087^{\pm.013}$
7-w 3-s	TMR	$0.704^{\pm.009}$	$0.679^{\pm.010}$	$0.852^{\pm.008}$	$0.440^{\pm.009}$	$0.746^{\pm.008}$	$\mathbf{0.815}^{\pm.012}$	$0.052^{\pm.007}$
	MAML	$\mathbf{0.705}^{\pm.008}$	$\mathbf{0.694}^{\pm.010}$	$\mathbf{0.864}^{\pm.008}$	$\mathbf{0.450}^{\pm.009}$	$\mathbf{0.748}^{\pm.008}$	$0.794^{\pm.013}$	$\mathbf{0.042}^{\pm.008}$
10-w 3-s	TMR	$0.708^{\pm.007}$	$0.680^{\pm.008}$	$0.849^{\pm.007}$	$0.447^{\pm.007}$	$0.753^{\pm.007}$	$0.815^{\pm.010}$	$0.121^{\pm.012}$
	MAML	$\mathbf{0.729}^{\pm.007}$	$\mathbf{0.724}^{\pm.009}$	$\mathbf{0.862}^{\pm.007}$	$\mathbf{0.473}^{\pm.008}$	$\mathbf{0.769}^{\pm.007}$	$\mathbf{0.819}^{\pm.009}$	$\mathbf{0.078}^{\pm.009}$
5-w 5-s	TMR	$0.778^{\pm.010}$	$0.758^{\pm.012}$	$0.908^{\pm.007}$	$0.472^{\pm.011}$	$0.811^{\pm.010}$	$0.854^{\pm.012}$	$0.128^{\pm.017}$
	MAML	$\mathbf{0.785}^{\pm.011}$	$\mathbf{0.786}^{\pm.013}$	$\mathbf{0.916}^{\pm.008}$	$\mathbf{0.484}^{\pm.012}$	$\mathbf{0.818}^{\pm.010}$	$\mathbf{0.881}^{\pm.011}$	$\mathbf{0.080}^{\pm.010}$
7-w 5-s	TMR	$0.754^{\pm.009}$	$0.731^{\pm.011}$	$0.871^{\pm.008}$	$0.462^{\pm.010}$	$0.779^{\pm.010}$	$0.816^{\pm.009}$	$0.094^{\pm.014}$
	MAML	$0.754^{\pm.009}$	$\mathbf{0.738}^{\pm.010}$	$\mathbf{0.873}^{\pm.007}$	$\mathbf{0.465}^{\pm.009}$	$\mathbf{0.780}^{\pm.009}$	$\mathbf{0.827}^{\pm.010}$	$\mathbf{0.063}^{\pm.012}$
10-w 5-s	TMR	$0.753^{\pm.007}$	$0.728^{\pm.007}$	$0.881^{\pm.006}$	$0.487^{\pm.007}$	$0.795^{\pm.007}$	$0.816^{\pm.011}$	$0.072^{\pm.009}$
	MAML	$\mathbf{0.767}^{\pm.007}$	$\mathbf{0.751}^{\pm.008}$	$\mathbf{0.887}^{\pm.006}$	$\mathbf{0.504}^{\pm.007}$	$\mathbf{0.799}^{\pm.007}$	$\mathbf{0.832}^{\pm.009}$	$\mathbf{0.060}^{\pm.006}$

Table 2: Full 120-Class Evaluation: Prototypical Baseline vs MAML Global Adaptation

Model	mAP	Top-1	Top-5	Recall@1	Recall@5	VDR	FPR
Text-only (0-shot)	0.337	0.306	0.485	0.244	0.405	0.559	0.016
Proto 1-shot	$0.364^{\pm.014}$	$0.301^{\pm.023}$	$0.529^{\pm.018}$	$0.237^{\pm.020}$	$0.456^{\pm.015}$	$0.624^{\pm.047}$	$0.060^{\pm.016}$
MAML 1-shot	$0.399^{\pm.017}$	$0.297^{\pm.021}$	$0.578^{\pm.019}$	$0.228^{\pm.021}$	$0.523^{\pm.018}$	$0.667^{\pm.043}$	$0.030^{\pm.011}$
Proto 3-shot	$0.448^{\pm.011}$	$0.397^{\pm.020}$	$0.600^{\pm.010}$	$0.324^{\pm.020}$	$0.528^{\pm.010}$	$0.538^{\pm.035}$	$0.014^{\pm.009}$
MAML 3-shot	$0.483^{\pm.008}$	$0.391^{\pm.011}$	$0.657^{\pm.007}$	$0.312^{\pm.011}$	$0.599^{\pm.007}$	$0.645^{\pm.031}$	$0.009^{\pm.008}$
Proto 5-shot	$0.47^{\pm.006}$	$0.426^{\pm.007}$	$0.624^{\pm.005}$	$0.346^{\pm.007}$	$0.551^{\pm.009}$	$0.613^{\pm.044}$	$0.020^{\pm.003}$
MAML 5-shot	$0.51^{\pm.005}$	$0.420^{\pm.007}$	$0.688^{\pm.005}$	$0.341^{\pm.008}$	$0.630^{\pm.004}$	$0.731^{\pm.021}$	$0.012^{\pm.004}$
Proto 10-shot	$0.504^{\pm.006}$	$0.454^{\pm.007}$	$0.659^{\pm.005}$	$0.372^{\pm.007}$	$0.586^{\pm.007}$	$0.656^{\pm.026}$	$0.010^{\pm.003}$
MAML 10-shot	$0.544^{\pm.004}$	$0.460^{\pm.008}$	$0.726^{\pm.005}$	$0.373^{\pm.007}$	$0.671^{\pm.005}$	$0.698^{\pm.018}$	$0.008^{\pm.002}$
Proto 20-shot	$0.514^{\pm.005}$	$0.451^{\pm.010}$	$0.621^{\pm.004}$	$0.367^{\pm.010}$	$0.617^{\pm.003}$	$0.645^{\pm.031}$	$0.007^{\pm.002}$
MAML 20-shot	$0.573^{\pm.006}$	$0.475^{\pm.007}$	$0.766^{\pm.009}$	$0.384^{\pm.007}$	$0.712^{\pm.008}$	$0.688^{\pm.019}$	$0.007^{\pm.002}$
Proto 30-shot	$0.516^{\pm.007}$	$0.452^{\pm.009}$	$0.622^{\pm.005}$	$0.368^{\pm.010}$	$0.615^{\pm.005}$	$0.677^{\pm.005}$	$0.009^{\pm.001}$
MAML 30-shot	$0.585^{\pm.005}$	$0.489^{\pm.005}$	$0.766^{\pm.008}$	$0.396^{\pm.005}$	$0.717^{\pm.007}$	$0.731^{\pm.023}$	$0.010^{\pm.002}$
Fully Supervised (BCE)	**0.683**	**0.612**	**0.824**	**0.507**	**0.786**	0.634	**0.006**
Fully Supervised (BCE+Center+Inter)	0.674	0.608	0.810	0.502	0.772	**0.742**	0.013