# Supplementary Materials
# MuSACo: Multimodal Subject-Specific Selection and Adaptation for Expression Recognition with Co-training

Muhammad Osama Zeeshan[1], Natacha Gillet[2], Alessandro Lameiras Koerich[1],
Marco Pedersoli[1], Francois Bremond[4], Eric Granger[1]
[1] LIVIA, Dept. of Systems Engineering, ÉTS Montreal, Canada
[2] École Polytechnique, Palaiseau, France
[3] Centre INRIA d'Université Côte d'Azur, Sophia Antipolis, France
muhammad-osama.zeeshan.1@ens.etsmtl.ca, {eric.granger, marco.pedersoli, alessandro.koerich}@etsmtl.ca
natacha.gillet@polytechnique.edu, francois.bremond@inria.fr

## Contents



## 1. Additional Implementation Details

For visual, ResNet18 [4] network, physio network based on an LSTM-based 1D-CNN network, consisting of two convolutional layers, one LSTM layer, and one fully connected layer. For the audio modality, we employ the Pretrained Audio Neural Network (PANNs) [6] framework, specifically the CNN14 architecture, pretrained on AudioSet. Additionally, the expression head is constructed using two 2-MLP layers for each network. A batch size of 32 for the target subject with images resized to 100×100 resolution, and the model is trained for 20 epochs.

### 1.1. Algorithm

**(a) Selection of Subject-Specific Sources.** Algorithm 1 shows the subject-specific selection of source subjects with co-training. Given source subjects $\mathcal{S}$ and specific target subject $\mathbf{T}$, extracted features using visual $B_{\text{v}}$ and physiological $B_{\text{p}}$ encoders. Construct a similarity metric $z^v$ and $z^p$ by measuring CosineSimilarity between every source subject and target. To estimate the maximum score, we normalize the similarities, merge them, and sort them in descending order, followed by a threshold that selects the most relevant subjects from the target.

**(b) Training Protocol of MuSACo.** In algorithm 2, we show our training protocol of adapting to an unlabeled target subject. All the equation numbers referenced here correspond to those in the main paper.

### 1.2. Hyperparameters

**Training of backbones.** momentum=0.9, weight decay=5e-4, stochastic gradient descent (SGD) optimizer [9], learning rate=1e-4 with lr scheduler (eta_min=0.00002). **Weighting parameters.** In MuSACo,

**Algorithm 1** Subject-Specific Selection of Source Subjects

**Require:**
$\mathcal{S}$: labeled source subjects, $\mathbf{T}$: unlabeled target domain, $B_{\text{vis}}$: visual encoder, $B_{\text{phy}}$: physio encoder, $\tau_{ss}$: threshold to select relevant source subjects

**Initialize:** $\mathbf{z}^v \leftarrow []$, $\mathbf{z}^p \leftarrow []$
Decompose $\mathbf{T}$ into $\mathrm{T}^v$ and $\mathrm{T}^p$
# Extract target features
$\mathrm{X}_t^v \leftarrow B_{vis}(\mathrm{T}^v)$, $\mathrm{X}_t^p \leftarrow B_{phy}(\mathrm{T}^p)$
**for** each domain $\mathbf{S}_i \in \mathcal{S}$ **do**
    Decompose $\mathbf{S}_i$ into $\mathrm{S}^v$ and $\mathrm{S}^p$
    # Extract source features
    $\mathrm{X}_s^v \leftarrow B_{vis}(\mathrm{S}^v)$, $\mathrm{X}_s^p \leftarrow B_{phy}(\mathrm{S}^p)$
    # Compute similarities
    $\mathrm{z}_i^v \leftarrow \cos(\mathrm{X}_s^v, \mathrm{X}_t^v)$
    $\mathrm{z}_i^p \leftarrow \cos(\mathrm{X}_s^p, \mathrm{X}_t^p)$
    # Append similarities
    $\mathbf{z}^v \leftarrow \mathbf{z}^v \cup \mathrm{z}_i^v$, $\mathbf{z}^p \leftarrow \mathbf{z}^p \cup \mathrm{z}_i^p$
**end for**
$\mathbf{z}^v \leftarrow \mathrm{norm}(\mathbf{z}^v)$, $\mathbf{z}^p \leftarrow \mathrm{norm}(\mathbf{z}^p)$
Merge similarities: $\mathbf{z} \leftarrow \mathrm{merge}(\mathbf{z}^v, \mathbf{z}^p)$
Sort $\mathbf{z}$ in descending order
Select relevant sources : $\widetilde{\mathcal{S}} \leftarrow \tau_{ss}(\mathbf{z})$

we give different weights to different loss functions. For the contribution of class-agnostic loss $\gamma = 0.5$, and for class-aware loss $\alpha = 0.1$. **Disentanglement.** Knife is very sensitive to hyperparameters; we have explored several parameters to make it work with the expression recognition task. The most critical parameters that is selected for our experiments are: zd_dim=1024, zc_dim=77, hidden_state=512, layers=3, nb_mixture=10, with learning-rate=0.01.

## 2. Detail on Training the Source Backbones

Training of the source backbones (visual and physiological) involves disentangling the identity-related features from the expression task. Our method is inspired by KNIFE[8], a fully differentiable entropy estimator, which we adapted for disentanglement in a multi-modal framework. The KNIFE estimator optimizes the backbones (visual and physiological) by decoupling non-task-related information through individual modality-specific gradient-based optimization. We leverage the KNIFE estimator to minimize the mutual information between modalities in eliminating identity features and enhancing the disentanglement of task-relevant features across distributions. We first estimate the marginal entropy of embedding $\mathbf{h}_m^{\mathbf{s}}$ as,

$$H(\mathbf{h}_m^{\mathbf{s}}) = -\mathbb{E}[MMp(\mathbf{h}_m^{\mathbf{s}})] \quad (1)$$

**Algorithm 2** Training protocol of MSACo

**Require:**
$\widetilde{\mathcal{S}}$: selected labeled source subjects
$\mathbf{T}$: unlabeled target domain

**for** epoch **do**
    Perform co-training on $\mathbf{T}$ to generate pseudo-labels (after every *n* epochs)
    **for** iteration **do**
        # Class-aware alignment
        Class-aware domain sampling of $\widetilde{\mathcal{S}}$ and $\mathbf{T}$
        Estimate intra-class discrepancy using (Eq: 8)
        Estimate inter-class discrepancy using (Eq: 9)
        Compute class-aware loss $\mathcal{L}_{aw}$ (Eq: 12)
        # Domain-agnostic alignment
        Domain-agnostic sampling of $\widetilde{\mathcal{S}}$ and $\mathbf{T}$
        Estimate domain agnostic using (Eq 10)
        Compute domain-agnostic loss $\mathcal{L}_{agn}$ (Eq: 11)
        # Modality alignment
        Perform feat concatenation of source subjects between modalities $\mathbf{h}_{\text{v}}^{\mathbf{s}}$ and $\mathbf{h}_{\text{p}}^{\mathbf{s}}$
        Perform feat concatenation of target subject between modalities $\mathbf{h}_{\text{v}}^{\mathbf{t}}$ and $\mathbf{h}_{\text{p}}^{\mathbf{t}}$
        Compute modality align loss $\mathcal{L}^{\mathbf{s}}$ using (Eq: 13)
        Compute modality align $\mathcal{L}_{\text{unsup}}^{\mathbf{t}}$ using (Eq: 14)
    **end for**
**end for**

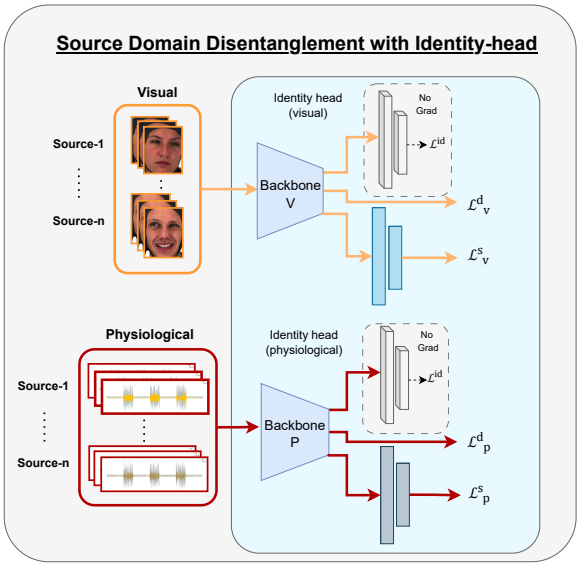


Figure 1. Training of disentanglement with knife-loss and identity-head.

where $\mathbb{E}$ is the expectation over the distribution $\mathbf{h}_m^{\mathbf{s}}$ and $p(.)$ is the probability density. To estimate the conditional en-

tropy between the features and identities, For each sample in the source subjects, we convert the identity labels $\acute{y}^{\mathbf{s}}_{m_j}$ into one-hot representations $\acute{Y}^{\mathbf{s}}_{m_j} \in R^k$. Then, conditional entropy is calculated as,

$$H(\mathbf{h}^{\mathbf{s}}_{m_j}|\acute{Y}^{s}_{m_j}) = -\mathbb{E}[\log p(\mathbf{h}^{\mathbf{s}}_{m_j}|\acute{Y}^{\mathbf{s}}_{m_j})] \quad (2)$$

where $p(.,.)$ is the conditional probability density between features $\mathbf{h}^{\mathbf{s}}_{m_j}$ and prediction $\acute{Y}^{\mathbf{s}}_{m_j}$. The total learning loss is calculated as:

$$\mathcal{L}^{d}_{m} = H(\mathbf{h}^{\mathbf{s}}_{m}) + H(\mathbf{h}^{\mathbf{s}}_{m_j}|\acute{Y}^{\mathbf{s}}_{m_j}). \quad (3)$$

By minimizing $\mathcal{L}^{d}_{m}$, the model is encouraged to decouple the information associated with $\acute{y}^{\mathbf{s}}_{m}$ from the feature embeddings $h^{\mathbf{s}}_{m}$. Furthermore, we introduce a fixed Identity-head $I_m$ head composed of two fully connected layers, as shown in Figure 1. It takes embeddings $\mathbf{h}^{\mathbf{s}}_{m}$ without gradient backpropagation. $I_m$ works as a regularizer in conjunction with the disentanglement loss to suppress non-discriminative identity-related information during target adaptation. The disentanglement loss decouples identity-related information, while the $I_m$ helps to constrain redundant (non-task-specific) features, ensuring the model focuses on learning relevant expression-related representations.

# 3. Datasets

## 3.1. Biovid Heat Pain

[12] dataset consists of 87 subjects captured in a controlled environment. It consists of five classes, including: *No-Pain*, *PA-1*, *PA-2*, *PA-3*, and *PA-4*. Every individual recorded 20 videos per class, corresponding to 100 videos per subject. We follow the same protocol as [13], where it eliminates the initial 2 seconds from the video, which does not show any spike indicating pain. In our experiment, we follow [14] to consider 77 subjects as sources and 10 subjects as targets: Sub-1 (081014_w_27), Sub-2 (101609_m_36), Sub-3 (112009_w_43), Sub-4 (091809_w_43), Sub-5 (071309_w_21), Sub-6 (073114_m_25), Sub-7 (080314_w_25), Sub-8 (073109_w_28), Sub-9 (100909_w_65), Sub-10 (081609_w_40). The label distribution statistics for the test set of Biovid were shown in Tab. 1. Each subject has nearly balanced distributions across the five classes (neutral + 4 pain levels), confirming the dataset is overall class-balanced.

## 3.2. StressID

[1] dataset comprised 65 (47 men and 18 women) subjects taken in a lab-controlled environment. Each participant was exposed to 11 tasks, grouped into 4 categories: watching emotional videos, breathing, interactive tasks, and relaxation. These tasks were captured using three different modalities (visual, physiological, and audio). Due to the unavailability of certain modalities in 11 of the participants, we only consider 54 individuals, which includes all the modalities. In our experiments, we take 44 subjects as sources and 10 subjects as targets. The target subject selection follows a 70% (men) and 30% (women) ratio, reflecting the higher number of male participants in the dataset. For each participant, there are up to 11 task-specific videos, from which we extracted frames at 1 fps for per-image expression classification. Tab. 2 shows the selected target subject demographics (ID, Gender) and the total number of samples. Tab. 3 shows the label distribution of StressID datasets. It exhibits a noticeable imbalance between stress and no-stress samples across subjects. Only two subjects, Sub-1 and Sub-8, are somewhat balanced, while others are skewed.

## 3.3. Behavioural Ambivalence/Hesitancy (BAH)

The Behavioural Ambivalence/Hesitancy (BAH) [3] dataset is the first multimodal resource designed for subject-specific recognition of ambivalence and hesitancy (A/H) in real-world settings. It consists of 1,118 videos (8.26 hours, including 1.5 hours of A/H) from 224 participants across 9 Canadian provinces, covering diverse ages and ethnicities. Participants answered 7 questions, some explicitly eliciting A/H, while being recorded via webcam and microphone, providing multimodal signals such as facial expressions, vocal cues, and transcripts. Expert annotations mark A/H occurrences at both frame- and video-level, with aligned face crops, transcripts, and participant metadata also included. Unlike controlled lab datasets (e.g., BioVid, StressID), BAH captures natural, in-the-wild behaviours, making it uniquely challenging and valuable for multimodal domain adaptation. In our experiments, we use 143 training data from BAH as subjects to consider as source domains and use 5 subjects from the test-set of BAH as targets: Sub-1 (82711), Sub-2 (82585), Sub-3 (82683), Sub-4 (82708), and Sub-5 (82632), enabling a realistic evaluation of `MuSACo` under noisy, heterogeneous conditions. The label distribution statistics for the test set of Stress-ID were shown in Tab. 4 It also shows a significant imbalance in ambivalence vs. non-ambivalence across most subjects, with only Sub-1 (82711) and Sub-8 (82683) being closer to balanced.

# 4. Ablation

## 4.1. Impact of Weighting Hyperparameters.

We performed a sequential weight sensitivity analysis for the key loss terms, $\mathcal{L}^{t}_{unsup}$, $\mathcal{L}_{aw}$, and $\mathcal{L}_{agn}$ shown in Tab. 5). The weight sensitivity analysis was performed on three representative subjects from the BioVid dataset: a young woman (081014_w_27), a young man (073114_m_25), and an older woman (100909_w_65). The reported results were the average of all three subjects, and each experiment was

| Subjects | ID | NA | PA1 | PA2 | PA3 | PA4 | Total |
|---|---|---|---|---|---|---|---|
| Sub-1 | 081014_w_27 | 287 | 285 | 311 | 305 | 312 | 1500 |
| Sub-2 | 101609_m_36 | 325 | 278 | 323 | 274 | 300 | 1500 |
| Sub-3 | 112009_w_43 | 304 | 282 | 296 | 310 | 308 | 1500 |
| Sub-4 | 091809_w_43 | 265 | 316 | 295 | 298 | 302 | 1476 |
| Sub-5 | 071309_w_21 | 299 | 297 | 295 | 304 | 297 | 1492 |
| Sub-6 | 073114_m_25 | 295 | 310 | 320 | 283 | 292 | 1500 |
| Sub-7 | 080314_w_25 | 311 | 318 | 288 | 276 | 307 | 1500 |
| Sub-8 | 073109_w_28 | 314 | 300 | 300 | 272 | 303 | 1489 |
| Sub-9 | 100909_w_65 | 298 | 298 | 313 | 279 | 312 | 1500 |
| Sub-10 | 081609_w_40 | 300 | 297 | 313 | 286 | 303 | 1499 |

Table 1. Label distribution for BioVid dataset across target subjects test-set.

| Subjects | ID | Gender | NSV | SV | No. frames |
|---|---|---|---|---|---|
| Sub-1 | kycf | Men | 5 | 6 | 1021 |
| Sub-2 | uymz | Men | 8 | 3 | 880 |
| Sub-3 | h8s1 | Men | 4 | 7 | 957 |
| Sub-4 | ctzy | Men | 2 | 9 | 1041 |
| Sub-5 | p9i3 | Men | 5 | 6 | 1041 |
| Sub-6 | 7h5u | Men | 5 | 5 | 917 |
| Sub-7 | g7r2 | Men | 8 | 3 | 1041 |
| Sub-8 | b9w0 | Women | 4 | 7 | 1026 |
| Sub-9 | r3zm | Women | 3 | 8 | 909 |
| Sub-10 | x1q3 | Women | 2 | 9 | 1001 |

Table 2. StressID target subject demographics and total number of samples (NSV: No Stress Videos, SV: Stress Videos)

| Subjects | ID | No-Stress | Stress | Total |
|---|---|---|---|---|
| Sub-1 | kycf | 118 | 86 | 204 |
| Sub-2 | uymz | 148 | 28 | 176 |
| Sub-3 | h8s1 | 128 | 83 | 211 |
| Sub-4 | ctzy | 60 | 148 | 208 |
| Sub-5 | p9i3 | 125 | 83 | 208 |
| Sub-6 | 7h5u | 127 | 57 | 184 |
| Sub-7 | g7r2 | 175 | 33 | 208 |
| Sub-8 | b9w0 | 105 | 100 | 205 |
| Sub-9 | r3zm | 63 | 120 | 183 |
| Sub-10 | x1q3 | 59 | 142 | 201 |

Table 3. Label distribution for StressID dataset across target subjects test-set.

run for 10 epochs. These subjects were chosen to reflect the diversity of the dataset across age and gender groups. First, we varied the weight of the target PL loss while keeping the others fixed at 1. Using the best result, we tuned the class-aware alignment loss, followed by the class-agnostic loss,

| Subjects | ID | No-Amb | Amb | Total |
|---|---|---|---|---|
| Sub-1 | 82711 | 433 | 874 | 1307 |
| Sub-2 | 82687 | 319 | 207 | 526 |
| Sub-3 | 82585 | 575 | 85 | 660 |
| Sub-4 | 82592 | 737 | 93 | 830 |
| Sub-5 | 82598 | 1199 | 189 | 1388 |
| Sub-6 | 82632 | 854 | 243 | 1097 |
| Sub-7 | 82681 | 354 | 207 | 561 |
| Sub-8 | 82683 | 445 | 456 | 901 |
| Sub-9 | 82708 | 249 | 104 | 353 |
| Sub-10 | 82714 | 231 | 234 | 465 |

Table 4. Label distribution for BAH dataset across target subjects test-set.

each time fixing the previously selected best weights. This sequential tuning strategy allowed us to identify the most effective configuration while minimizing computation cost. The best configuration obtained was $\gamma_t = 1$, $\alpha = 0.1$, and $\beta = 0.5$. All experiments in the paper were conducted using this configuration to ensure consistency across results.

| Weights | Accuracy | | |
|---|---|---|---|
| | $\mathcal{L}^t_{unsup}$ | $\mathcal{L}_{aw}$ | $\mathcal{L}_{agn}$ |
| **0.01** | 38.6 | 40.7 | 40.9 |
| **0.05** | 40.5 | 40.7 | 43.8 |
| **0.1*** | 40.3 | **43.9** | 41.4 |
| **0.3** | 39.9 | 39.9 | 40.1 |
| **0.5*** | 38.7 | 40.3 | **44.1** |
| **0.7** | 39.9 | 39.9 | 39.9 |
| **0.9** | 40.2 | 41.4 | 43.9 |
| **1*** | **40.7** | 40.7 | 40.9 |
| **2** | 40.6 | 39.6 | 41.9 |
| **5** | 40.2 | 38.8 | 22.0 |

Table 5. Ablation study on loss weights. Each loss weight is varied while keeping the previously selected optimal weights fixed.

## 4.2. Source Subject Selection Analysis

Tab. 6 shows the number of selected source subjects for each of the 10 target subjects for the Biovid dataset for various threshold ($\tau_{ss}$) values. As $\tau_{ss}$ increases, fewer source subjects are selected, reducing potential noise from irrelevant domains. Notably, the accuracy improves when transitioning from $\tau = 0$ to moderate thresholds (e.g., $\tau_{ss} = 0.45$ or $0.55$), demonstrating that selective inclu-

| $\tau_{ss}$ | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Avg SS | Avg Acc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77.0 | 37.66 |
| 0.05 | 56 | 55 | 62 | 52 | 57 | 50 | 59 | 53 | 52 | 65 | 56.1 | 37.59 |
| 0.10 | 53 | 50 | 57 | 52 | 53 | 48 | 57 | 52 | 49 | 63 | 53.4 | 38.02 |
| 0.15 | 50 | 47 | 56 | 49 | 49 | 45 | 52 | 50 | 44 | 61 | 50.3 | 33.92 |
| 0.20 | 48 | 44 | 50 | 47 | 45 | 43 | 48 | 44 | 40 | 54 | 46.3 | 36.51 |
| 0.25 | 43 | 39 | 44 | 44 | 45 | 43 | 41 | 40 | 35 | 51 | 42.5 | 36.70 |
| 0.30 | 39 | 35 | 40 | 37 | 43 | 41 | 37 | 38 | 33 | 49 | 39.2 | 38.90 |
| 0.35 | 35 | 28 | 34 | 35 | 38 | 36 | 32 | 36 | 29 | 44 | 34.7 | 35.52 |
| 0.40 | 31 | 22 | 28 | 32 | 35 | 33 | 30 | 28 | 27 | 39 | 30.5 | 37.27 |
| 0.45 | 31 | 20 | 26 | 26 | 31 | 30 | 25 | 20 | 21 | 35 | 26.5 | 38.39 |
| 0.50 | 26 | 16 | 24 | 26 | 28 | 26 | 21 | 16 | 18 | 34 | 23.5 | 38.69 |
| **0.55** | **20** | **23** | **23** | **24** | **22** | **23** | **16** | **13** | **15** | **30** | **20.9** | **39.89** |
| 0.60 | 16 | 15 | 21 | 20 | 15 | 18 | 16 | 10 | 12 | 27 | 17.0 | 38.29 |
| 0.65 | 12 | 13 | 20 | 19 | 15 | 16 | 11 | 11 | 10 | 22 | 14.9 | 37.62 |
| 0.70 | 9 | 8 | 15 | 15 | 12 | 15 | 7 | 6 | 9 | 21 | 11.7 | 37.48 |
| 0.75 | 7 | 6 | 13 | 12 | 10 | 13 | 5 | 5 | 8 | 20 | 9.9 | 37.42 |
| 0.80 | 5 | 5 | 9 | 7 | 6 | 11 | 3 | 4 | 8 | 16 | 7.4 | 31.31 |
| 0.85 | 4 | 4 | 6 | 7 | 5 | 9 | 3 | 3 | 5 | 13 | 5.9 | 30.27 |
| 0.90 | 2 | 2 | 6 | 6 | 4 | 7 | 3 | 3 | 3 | 10 | 4.6 | 27.93 |
| 0.95 | 2 | 2 | 4 | 3 | 3 | 3 | 2 | 2 | 3 | 5 | 2.9 | 28.53 |
| 1.00 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2.0 | 29.12 |

Table 6. Number of selected subjects (SS) for each target subject across different $\tau_{ss}$ thresholds. Lower $\tau_{ss}$ values indicate fewer selected sources. We also report the average number of selected subjects and the corresponding average accuracy.

sion of source data benefits generalization. The best average accuracy of 39.89% is achieved with $\tau_{ss} = 0.55$, using approximately 21 subjects on average. To assess the robustness of the source subject selection threshold ($\tau_{ss}$), we conducted experiments on the Behavioural Ambivalence/Hesitancy (BAH) [3] dataset, which differs substantially from lab-controlled datasets (BioVid and Stress) as it consists of self-recorded, in-the-wild videos. Table 7 shows results across thresholds from 0.45 to 0.65. The analysis reveals that accuracy remains stable in the 0.45–0.55 range, with the best performance achieved at $\tau_{ss} = 0.55$. Higher thresholds reduce the number of selected sources and lead to performance degradation due to over-pruning. These findings confirm that $\tau_{ss}$ is not highly sensitive to fine-tuning and that a default value around 0.55 provides a reliable trade-off between accuracy and source diversity across datasets.

### 4.3. Ablation on Target PLs Threshold ($\tau_{pl}$).

Tab. 8, and found that $\tau_0 = 0.95$ provided the best trade-off between label quality and sample quantity. Rather than fixing this value, we employ a confidence-annealing schedule, where $\tau_{pl}$ is decreased by $0.01$ every $N$ epochs, allowing more target samples to be included as the model stabilizes. This ensures early training benefits from high-confidence labels, while later stages incorporate a broader set of samples. To further mitigate the exclusion of useful but uncertain samples, we introduce a domain-agnostic loss that allows learning from low-confidence examples as well. Target ground-truth labels were only used here to evaluate pseudo-label quality.

### 4.4. Impact of Disentanglement

conducted separately for visual and physiological modalities on the BioVid dataset, using 77 source and 10 target subjects. Results (Fig. 2, Tab. 9) show that incorporating the KNIFE-based disentanglement improves performance for both modalities. Without disentanglement, iden-

| $\tau_{ss}$ | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Avg SS | Avg Acc |
|---|---|---|---|---|---|---|---|
| 0.45 | 44 | 33 | 39 | 41 | 39 | 39.2 | 67.1 |
| 0.50 | 42 | 31 | 36 | 40 | 36 | 37.0 | 68.2 |
| **0.55** | **41** | **30** | **35** | **38** | **34** | **35.6** | **68.9** |
| 0.60 | 39 | 28 | 33 | 37 | 33 | 34.0 | 66.6 |
| 0.65 | 38 | 27 | 30 | 36 | 30 | 32.2 | 62.5 |

Table 7. Impact of varying source subject selection threshold $\tau_{ss}$ on the BAH dataset.

| **Threshold ($\tau_{pl}$)** | **Accuracy** |
|---|---|
| 0.50 | 40.2 |
| 0.60 | 48.1 |
| 0.70 | 44.7 |
| 0.75 | 46.5 |
| 0.80 | 46.2 |
| 0.85 | 48.6 |
| 0.90 | 46.7 |
| **0.95** | **52.0** |

Table 8. Accuracy at different pseudo-label confidence thresholds ($\tau_{pl}$)

| **Exp-head** | **Id-head** | **Knife** | **Visual** | **Physio** |
|---|---|---|---|---|
| ✓ | | | 25.1 | 36.2 |
| ✓ | | ✓ | 26.3 | 37.0 |
| ✓ | ✓ | ✓ | **28.3** | **38.2** |

Table 9. Ablation on the disentanglement module for visual and physiological modalities for the Biovid dataset.

tity information interferes with expression classification, resulting in reduced accuracy. Adding the KNIFE-loss mitigates this, and introducing an identity head yields further improvements, confirming that explicitly removing identity bias enhances modality-specific expression features.

### 4.5. Computational Complexity

To analyze the time complexity, we evaluated all methods under identical conditions using a ResNet-18 backbone (11.7M parameters), a fixed batch size of 32, and the same GPU (NVIDIA A100-SXM4-40GB). The reported training time includes both source and target, and inference time with accuracy are summarized in Tab. 10. The significant training time of CAN [5] is due to its design, which processes each source domain separately and performs class-wise contrastive learning between each source and the target domain within every batch. While manageable for benchmark datasets like Office-31 (3 domains), this becomes prohibitively expensive when scaling to our subject-based adaptation setting, where the number of source domains averages over 30, resulting in 19,625 ms of target training

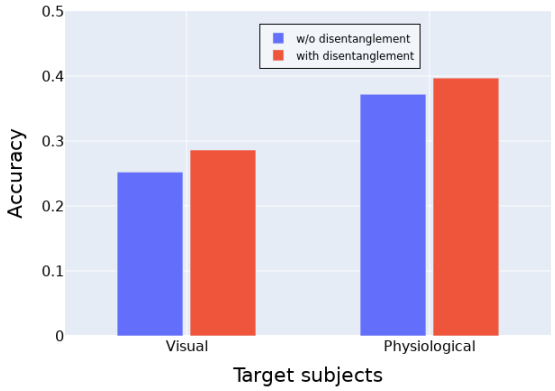


Figure 2. Impact of adding disentanglement.

time. In contrast, our method retrieves class-wise samples from all source domains and from the target domain, then performs a single contrastive operation per batch, significantly reducing overhead. It completes target training in just 426 ms (46× faster than CAN) while improving accuracy from 34.6% to 43.8%. Although CMSDA [10] and sub-based [14] adaptation methods are more efficient (271 ms and 234 ms), they fail to yield improvements in accuracy. Results show that our method offers the best trade-off between efficiency and accuracy, with inference time similar to other baselines.

| **Method** | **Src Train** | **Trg Train** | **Inf** | **Acc** |
|---|---|---|---|---|
| CAN [5] | 59 | 19,625 | 8.03 | 34.6 |
| Sub-based [14] | 59 | 234 | 7.36 | 34.6 |
| CMSDA [10] | – | 271 | 4.07 | 36.8 |
| Ours (all srcs) | 59 | 1895 | 7.82 | 39.3 |
| Ours (co-train) | **59** | **426** | **7.64** | **43.8** |

Table 10. Training and inference time (per batch) of our proposed and baseline methods. All times are shown in ms.

| $\mathcal{L}^s$ | $\mathcal{L}^t$ | $\mathcal{L}_{aw}$ | $\mathcal{L}_{agn}$ | **Avg Acc** |
|---|---|---|---|---|
| ✓ | | | | 32.1 |
| ✓ | ✓ | | | 36.5 |
| ✓ | ✓ | ✓ | | 38.0 |
| ✓ | ✓ | | ✓ | 37.2 |
| ✓ | ✓ | ✓ | ✓ | **40.7** |

Table 11. Impact of individual loss components under equally weighted hyperparameter.

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Avg F1 |
| Lower Bound | Visual-only | 63.1 | 24.6 | 61.2 | 37.8 | 58.3 | 50.8 | 43.2 | 36.6 | 46.9 | 54.8 | 47.7 |
| | Physio-only | 36.6 | 45.6 | 37.7 | 22.3 | 37.5 | 40.8 | 45.6 | 33.8 | 25.6 | 22.7 | 34.8 |
| | Fusion | 57.3 | 32.1 | 50.0 | 36.6 | 51.5 | 63.4 | 68.4 | 39.5 | 52.6 | 53.4 | 50.5 |
| MM-UDA (Blending) | DANN [2] | 61.2 | 45.3 | 55.6 | 51.3 | 53.6 | 56.6 | 56.6 | 47.6 | 62.3 | 38.6 | 52.9 |
| | CDAN [7] | 57.1 | 43.2 | 64.3 | 46.6 | 48.3 | 58.3 | 59.6 | 45.6 | 62.2 | 39.6 | 52.5 |
| | MMD [11] | 57.1 | 54.0 | 61.3 | 40.3 | 45.6 | 55.6 | 55.6 | 55.6 | 55.4 | 49.3 | 53.0 |
| | MuSACo (UDA) | 37.6 | 49.3 | 74.1 | 86.2 | 55.7 | 61.5 | 64.8 | 53.4 | 74.4 | 50.0 | 60.7 |
| MM-MSDA | CAN [5] | 65.3 | 45.4 | 54.6 | 42.3 | 51.2 | 58.6 | 66.6 | 54.2 | 59.3 | 34.6 | 53.2 |
| | Sub-based$_{\text{top-k}}$ [14] | 68.9 | 45.3 | 74.2 | 46.6 | 55.2 | 62.6 | 56.6 | 39.3 | 52.2 | 39.9 | 54.1 |
| | MuSACo (MSDA) | 76.1 | 59.6 | 69.3 | 53.3 | 58.6 | 67.2 | 75.2 | 48.3 | 74.6 | 40.2 | **62.2** |
| Upper Bound | Fine-tuning | 86.3 | 43.8 | 94.4 | 97.7 | 79.6 | 75.6 | 45.6 | 93.0 | 91.0 | 97.0 | 80.4 |

Table 12. Subject-wise F1 scores of StressID dataset across 10 target subjects.

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Avg F1 |
| Lower Bound | Visual-only | 29.8 | 26.1 | 9.1 | 28.1 | 19.3 | 21.6 | 25.7 | 15.3 | 16.5 | 16.1 | 20.8 |
| | Physio-only | 38.1 | 32.5 | 39.1 | 15.4 | 10.4 | 29.0 | 14.3 | 34.6 | 28.7 | 22.0 | 26.4 |
| | Fusion | 41.2 | 27.2 | 24.9 | 32.2 | 29.5 | 25.8 | 34.7 | 24.3 | 22.3 | 18.9 | 28.1 |
| MM-UDA (Blending) | DANN [2] | 32.5 | 25.6 | 29.5 | 33.6 | 28.3 | 25.1 | 29.6 | 24.2 | 25.2 | 20.2 | 27.4 |
| | CDAN [7] | 28.5 | 24.3 | 25.5 | 29.2 | 25.2 | 24.2 | 28.6 | 23.4 | 23.1 | 20.5 | 25.3 |
| | MMD [11] | 28.3 | 20.7 | 29.6 | 32.4 | 28.4 | 29.5 | 32.1 | 28.6 | 25.4 | 26.5 | 28.2 |
| | MuSACo (UDA) | 35.6 | 21.7 | 33.3 | 36.2 | 28.7 | 16.0 | 31.2 | 29.4 | 35.4 | 17.6 | 28.5 |
| MM-MSDA | CAN [5] | 34.6 | 30.3 | 22.3 | 24.6 | 27.9 | 26.3 | 28.6 | 26.6 | 33.2 | 18.1 | 27.3 |
| | Sub-based$_{\text{top-k}}$ [14] | 34.6 | 37.6 | 22.4 | 20.5 | 34.5 | 24.6 | 36.8 | 24.3 | 30.6 | 17.6 | 28.4 |
| | CMSDA [10] | 31.2 | 32.5 | 21.2 | 31.2 | 32.3 | 30.1 | 30.2 | 34.1 | 38.9 | 18.5 | 30.0 |
| | MuSACo (MSDA) | 41.6 | 33.2 | 49.6 | 27.8 | 25.6 | 36.5 | 29.4 | 35.4 | 49.0 | 29.2 | **35.7** |
| Upper Bound | Fine-tuning | 81.2 | 74.3 | 67.6 | 68.2 | 70.3 | 61.4 | 74.2 | 70.6 | 74.0 | 49.5 | 69.1 |

Table 13. Subject-wise F1 scores of BioVid dataset on 10 target subjects.

### 4.6. Impact of Difference Loss Components with Equal Weights

In addition to the tuned-weight ablation reported in the main paper, we also conducted a diagnostic experiment where all loss weights were fixed to 1. This isolates the relative impact of each component without any hyperparameter tuning. Table 11 shows that the full model under equal weights reached 40.7%, which is lower than the tuned configuration (43.8%). Importantly, the incremental trend remains consistent across both setups: $\mathcal{L}^t$, $\mathcal{L}_{aw}$, and $\mathcal{L}_{agn}$ each contribute positively to performance. This confirms that the effectiveness of MuSACo does not hinge on carefully tuned weights, but rather that appropriate weighting further amplifies the gains. Together, these results underline the robustness and general applicability of the proposed loss formulation.

### 4.7. Target Pseudo-label Progression During Training

We evaluate the reliability of co-training by tracking pseudo-label (PL) accuracy during training. Results are reported for two challenging StressID subjects: ctzy (Sub-4) and x1q3 (Sub-10), which exhibit high label imbalance. To stabilize PL quality, we adopt a confidence-annealing schedule, initializing $\tau_{pl} = 0.95$ and gradually reducing it by 0.01 every $N$ epochs (e.g., 0.94 at Ep-4, 0.93 at Ep-8). As shown in Fig. 3, PL accuracy steadily improves across epochs, demonstrating the effectiveness of iterative refinement under this schedule. For Sub-4, PL accuracy improves from 0.976 at initialization to 0.996 after 8 epochs, while for Sub-10, which is highly imbalanced, accuracy increases from 0.602 to 0.666. These trends indicate that co-training with confidence annealing reliably enhances PL quality, even under challenging conditions.

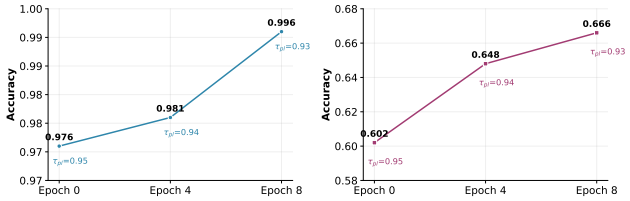


Figure 3. Target pseudo-labels accuracy during training.

| Settings | Methods | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Avg F1 |
|---|---|---|---|---|---|---|---|
| Lower Bound | Visual-only | 25.2 | 46.2 | 35.2 | 41.3 | 43.2 | 38.2 |
| | Audio-only | 27.4 | 34.4 | 34.1 | 40.3 | 44.1 | 36.1 |
| | Fusion | 26.3 | 44.5 | 34.7 | 41.9 | 44.6 | 38.4 |
| MM-UDA | MMD [11] | 28.2 | 44.5 | 39.3 | 38.7 | 43.2 | 38.8 |
| | MuSACo (UDA) | 26.1 | 46.1 | 35.6 | 42.2 | 43.5 | 38.7 |
| MM-MSDA | Sub-based$_{\text{top-k}}$ [14] | 24.6 | 45.4 | 36.8 | 42.7 | 44.9 | 38.9 |
| | MuSACo (MSDA) | **28.2** | **48.6** | **39.4** | **44.2** | **45.4** | **41.2** |

Table 14. Subject-wise F1 scores on the BAH dataset across five target subjects. MuSACo MSDA consistently outperforms baselines and other UDA/MSDA methods.

# 5. Subject-wise F1 Scores

## 5.1. Stress-ID

Table 12 presents subject-wise F1 scores on the StressID dataset. While the dataset has a binary label distribution (stress vs. no-stress), clear variability emerges across subjects. For example, lower-bound fusion consistently yields higher F1 than either modality alone, showing the complementary nature of visual and physiological signals. However, without adaptation, predictions remain unbalanced across subjects, with several experiencing sharp drops (e.g., Sub-2, Sub-4). UDA methods improve F1 moderately on some subjects, but results remain inconsistent. In contrast, MuSACo (MSDA) achieves the best overall performance, boosting recall on harder subjects (e.g., Sub-9) and producing the highest average F1 across all targets. These findings underscore the importance of F1 as a complementary metric to accuracy, particularly for stress recognition, where subject-specific imbalance strongly affects model robustness.

## 5.2. Biovid

Tab. 13 reports subject-wise F1 scores on the BioVid dataset. Our analysis shows that MuSACo particularly struggles on subjects where the Physio-only baseline itself is weak (e.g., Sub-4, Sub-5, Sub-7). This suggests that physiological variability across individuals is the key limiting factor: when physio cues are unreliable, visual signals alone are too subtle to compensate, and fusion inherits this weakness. These subject-specific outcomes highlight an inherent challenge of BioVid, where the modality that is usually strongest (physio) can be inconsistent for some individuals, constraining overall adaptation performance. Nevertheless, MuSACo achieves consistent performance across all subjects and outperforms competing methods on 6 out of 10 targets, leading to the best overall F1 score.

## 5.3. BAH

Tab. 14 reports subject-wise F1 scores on the BAH dataset. While the visual modality is generally stronger than audio, naive fusion provides slight gains by leveraging complementary information across modalities. MuSACo (MSDA) achieves the best overall performance, consistently surpassing both lower-bound baselines and existing UDA/MSDA approaches, demonstrating the effectiveness of source subject selection and multimodal adaptation in more challenging, in-the-wild conditions.

# 6. Evaluation on Additional Backbone

In the main paper, all experiments are conducted using a ResNet-18 backbone to ensure fair comparison with prior work and maintain a computationally efficient setup. However, MuSACo is inherently backbone-agnostic and does not rely on architectural properties specific to convolutional networks. To verify that MuSACo extends beyond ResNet-based architectures, we additionally evaluate the method using a Vision Transformer (ViT). We use the same MuSACo protocol as before for all the experiments, including the same setup for baseline methods. **Biovid.** Tables 15 and 16 report the subject-wise Accuracy and F1 scores across 10 target subjects. MuSACo consistently surpasses the baseline UDA and MSDA methods, achieving the highest average Accuracy and F1. Notably, MuSACo maintains strong performance even in subjects where the baselines degrade, demonstrating stable target adaptation. **StressID.** Tables 17 and 18 present the subject-wise Accuracy and F1 results. MuSACo consistently outperforms the baseline UDA and MSDA methods across most subjects, achieving the highest average Accuracy and F1. In particular, MuSACo demonstrates strong improvements on challenging subjects such as Sub-3, Sub-4, and Sub-6, where existing approaches show significant performance drops. These results further validate that the effectiveness of MuSACo is preserved even when replacing ResNet-18 with a transformer-based backbone, confirming that the proposed method is robust and generalizes well across different architectural families.

# 7. MuSACo

MuSACo (inspired by the layered harmony of Musaca) integrates two synergistic modules for subject-specific adaptation: a co-training-based source subject selection module that identifies the most relevant sources using complementary cues from multiple modalities, and an adaptation mod-

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Avg Acc |
| Lower Bound | Visual-only | 37.8 | 27.7 | 23.3 | 22.6 | 27.5 | 21.6 | 37.0 | 20.0 | 22.0 | 24.0 | 26.4 |
| | Physio-only | 41.7 | 30.1 | 30.5 | 33.3 | 36.7 | 30.9 | 34.7 | 38.0 | 37.3 | 28.0 | 34.1 |
| | Fusion | 36.0 | 28.0 | 26.0 | 35.0 | 35.0 | 15.0 | 43.0 | 40.0 | 35.0 | 28.0 | 32.1 |
| MM-UDA | MMD [11] | 40.3 | 31.2 | 32.3 | 26.3 | 33.3 | 24.1 | 34.3 | 39.5 | 37.3 | 20.0 | 31.9 |
| | MuSACo (UDA) | 39.0 | 35.0 | 34.0 | 32.0 | 29.0 | 19.0 | 38.9 | 37.0 | 31.0 | 29.9 | 32.5 |
| MM-MSDA | Sub-based$_{\text{top-k}}$ [14] | 45.0 | 35.0 | 39.0 | 37.0 | 34.0 | 32.0 | 42.0 | 36.0 | 27.0 | 30.0 | 35.7 |
| | MuSACo (MSDA) | 42.0 | 38.0 | 51.0 | 30.0 | 28.0 | 39.0 | 43.0 | 35.0 | 43.0 | 33.0 | **38.2** |

Table 15. Subject-wise accuracy of the BioVid dataset using the ViT backbone across 10 target subjects.

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Sub-1 | Sub-2 | Sub-3 | Sub-4 | Sub-5 | Sub-6 | Sub-7 | Sub-8 | Sub-9 | Sub-10 | Avg F1 |
| Lower Bound | Visual-only | 35.6 | 27.5 | 22.5 | 12.5 | 19.7 | 19.6 | 35.3 | 15.6 | 14.9 | 23.2 | 22.6 |
| | Physio-only | 31.0 | 21.9 | 23.1 | 24.5 | 27.9 | 23.6 | 23.9 | 27.3 | 28.2 | 19.1 | 25.0 |
| | Fusion | 23.9 | 27.1 | 22.8 | 22.6 | 33.5 | 9.5 | 32.9 | 34.6 | 29.1 | 26.1 | 26.2 |
| MM-UDA | MMD [11] | 31.7 | 33.1 | 28.6 | 20.5 | 33.8 | 23.5 | 37.7 | 35.6 | 29.3 | 18.1 | 29.2 |
| | MuSACo (UDA) | 33.6 | 34.6 | 25.9 | 26.9 | 25.9 | 18.4 | 38.4 | 33.8 | 31.3 | 29.8 | 29.9 |
| MM-MSDA | Sub-based$_{\text{top-k}}$ [14] | 36.9 | 27.2 | 34.6 | 32.9 | 20.6 | 27.4 | 34.5 | 28.0 | 13.1 | 22.6 | 27.8 |
| | MuSACo (MSDA) | 33.9 | 27.6 | 38.7 | 26.4 | 24.9 | 33.9 | 41.5 | 30.1 | 37.6 | 29.8 | **32.4** |

Table 16. Subject-wise F1 scores of the BioVid dataset using the ViT backbone across 10 target subjects.

ule that aligns source and target domains using class-aware and class-agnostic losses. Like Musaca's layered composition—where distinct elements work in harmony, MuSACo leverages modality-specific strengths to guide pseudo-label generation and fuse information effectively. This enables robust, personalized adaptation for each target subject while maintaining consistency across modalities.

## References


[1] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma Ismailova, Massimiliano Todisco, Maria A Zuluaga, et al. Stressid: a multimodal dataset for stress identification. *Advances in Neural Information Processing Systems*, 36:29798–29811, 2023. 3

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 7

[3] Manuela González-González, Soufiane Belharbi, Muhammad Osama Zeeshan, Masoumeh Sharafi, Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro Lameiras Koerich, Simon L Bacon, and Eric Granger. Bah dataset for ambivalence/hesitancy recognition in videos for behavioural change. *arXiv preprint arXiv:2505.19328*, 2025. 3, 5

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, Piscataway, USA, 2016. IEEE. 1

[5] Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE TPAMI*, 44(4):1793–1804, 2020. 6, 7

[6] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 1

[7] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 31, 2018. 7

[8] Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pages 17691–17715. PMLR, PMLR, 2022. 2

[9] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 1

[10] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization. *arXiv preprint arXiv:2106.16093*, 2021. 6, 7

[11] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pages 2263–2291, 2013. 7, 8, 9, 10

[12] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recog-

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Sub-1** | **Sub-2** | **Sub-3** | **Sub-4** | **Sub-5** | **Sub-6** | **Sub-7** | **Sub-8** | **Sub-9** | **Sub-10** | **Avg Acc** |
| Lower Bound | Visual-only | 57.0 | 18.3 | 41.7 | 68.3 | 62.4 | 47.5 | 84.2 | 47.0 | 65.5 | 69.7 | 56.2 |
| | Physio-only | 57.8 | 84.1 | 60.7 | 28.8 | 60.1 | 69.0 | 84.1 | 51.2 | 34.4 | 29.4 | 56.0 |
| | Fusion | 48.3 | 14.0 | 49.3 | 78.3 | 66.0 | 49.2 | 74.0 | 51.3 | 64.2 | 73.5 | 56.8 |
| MM-UDA | MMD [11] | 53.0 | 23.0 | 36.4 | 82.0 | 69.0 | 47.6 | 78.0 | 46.6 | 77.0 | 66.0 | 57.9 |
| | `MuSACo (UDA)` | 64.0 | 20.0 | 38.0 | 79.0 | 65.0 | 50.0 | 82.0 | 50.0 | 66.0 | 68.0 | 58.2 |
| MM-MSDA | Sub-based$_{\text{top-k}}$ [14] | 58.0 | 15.0 | 50.0 | 72.0 | 62.0 | 82.0 | 94.0 | 82.0 | 66.0 | 67.0 | 64.8 |
| | `MuSACo (MSDA)` | 38.0 | 38.0 | 87.0 | 100.0 | 67.0 | 67.0 | 85.0 | 82.0 | 66.0 | 68.0 | **69.8** |

Table 17. Subject-wise accuracy of the StressID dataset across 10 target subjects using the ViT backbone.

| Settings | Methods | Target Subjects | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Sub-1** | **Sub-2** | **Sub-3** | **Sub-4** | **Sub-5** | **Sub-6** | **Sub-7** | **Sub-8** | **Sub-9** | **Sub-10** | **Avg F1** |
| Lower Bound | Visual-only | 55.2 | 17.3 | 32.5 | 67.9 | 46.1 | 46.3 | 76.4 | 32.9 | 39.6 | 41.1 | 45.5 |
| | Physio-only | 36.6 | 45.7 | 37.8 | 22.4 | 37.5 | 40.8 | 45.7 | 33.9 | 25.6 | 22.7 | 34.9 |
| | Fusion | 47.0 | 13.0 | 36.4 | 77.4 | 53.4 | 47.1 | 64.2 | 42.8 | 39.6 | 42.3 | 46.3 |
| MM-UDA | MMD [11] | 51.6 | 22.6 | 27.5 | 72.6 | 55.2 | 42.1 | 72.5 | 44.6 | 44.0 | 39.5 | 47.2 |
| | `MuSACo (UDA)` | 62.6 | 19.3 | 28.5 | 77.4 | 49.3 | 48.5 | 72.8 | 34.8 | 39.5 | 40.2 | 47.3 |
| MM-MSDA | Sub-based$_{\text{top-k}}$ [14] | 56.0 | 13.8 | 46.5 | 71.0 | 38.0 | 80.1 | 87.3 | 81.2 | 39.5 | 39.9 | 55.3 |
| | `MuSACo (MSDA)` | 28.5 | 36.4 | 85.2 | 98.8 | 66.3 | 66.2 | 45.7 | 81.2 | 39.5 | 40.2 | **58.8** |

Table 18. Subject-wise F1 scores of the StressID dataset across 10 target subjects using the ViT backbone.

nition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE, 2013. 3

[13] Philipp Werner, Ayoub Al-Hamadi, and Steffen Walter. Analysis of facial expressiveness during experimentally induced heat pain. In *2017 Seventh international conference on affective computing and intelligent interaction workshops and demos (ACIIW)*, pages 176–180. IEEE, 2017. 3

[14] Muhammad Osama Zeeshan, Muhammad Haseeb Aslam, Soufiane Belharbi, Alessandro Lameiras Koerich, Marco Pedersoli, Simon Bacon, and Eric Granger. Subject-based domain adaptation for facial expression recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 3, 6, 7, 8, 9, 10