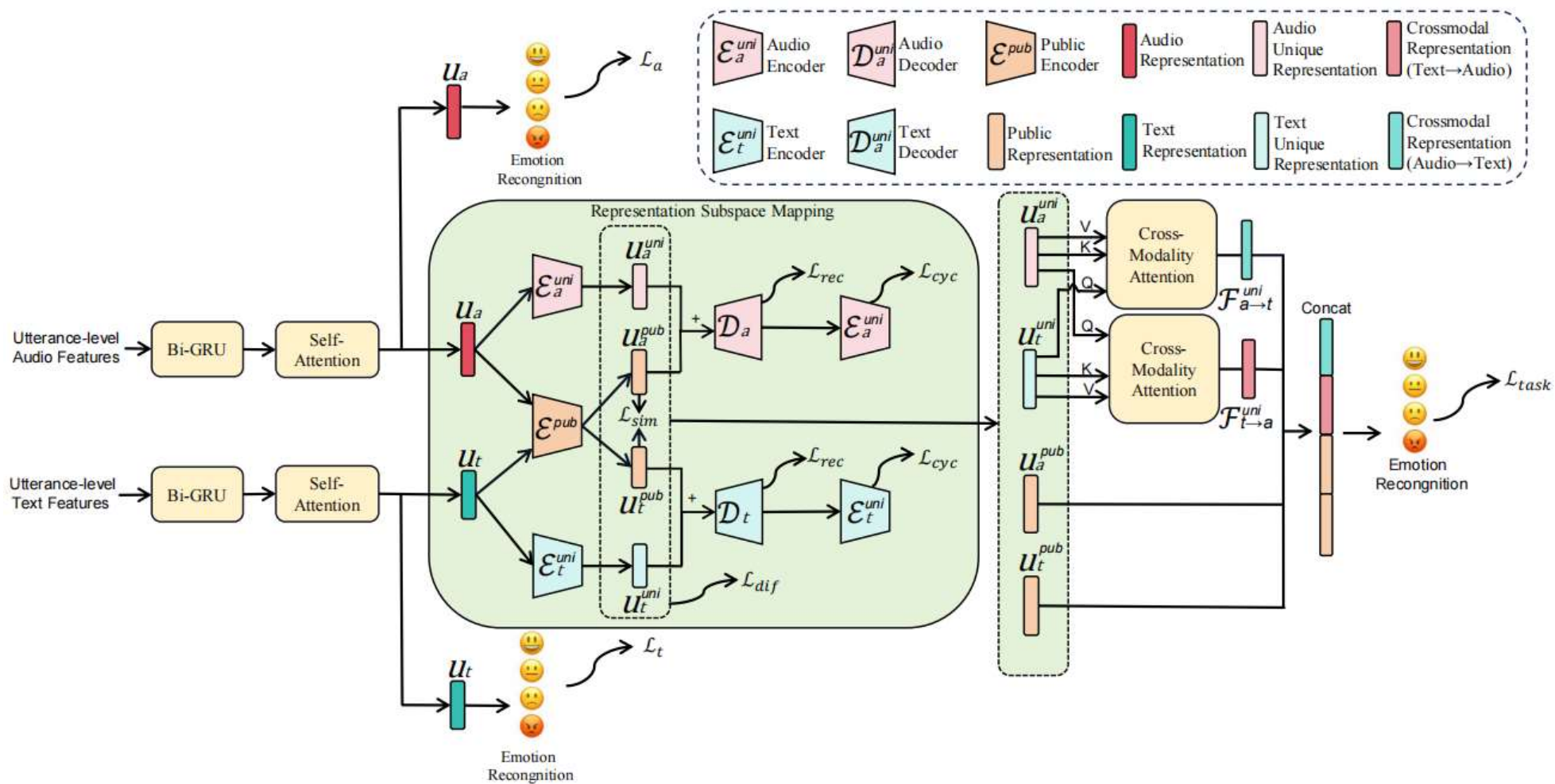


# **Integrating Representation Subspace Mapping with Unimodal Auxiliary Loss for Attention-based Multimodal Emotion Recognition**

Reported by Xulong Du

# Method



# Experiments

| Dataset    | Modality                      | Happy(%)     | Sad(%)       | Neutral(%)   | Angry(%)     | Average(%)   |
|------------|-------------------------------|--------------|--------------|--------------|--------------|--------------|
| IEMOCAP    | A                             | 57.11        | 64.49        | 66.15        | <b>91.18</b> | 69.73        |
|            | T                             | 90.74        | 88.16        | 72.39        | 71.18        | 80.62        |
|            | A+T (w/o $\mathcal{L}_{ua}$ ) | 86.00        | 87.76        | 66.15        | 86.47        | 81.10        |
|            | A+T                           | <b>94.01</b> | <b>88.30</b> | <b>73.40</b> | 88.00        | <b>85.92</b> |
| MSP-Improv | A                             | 84.55        | 72.73        | 67.21        | 51.79        | 69.07        |
|            | T                             | 87.27        | 69.32        | 72.13        | <b>90.18</b> | 82.00        |
|            | A+T (w/o $\mathcal{L}_{ua}$ ) | 87.27        | 77.27        | <b>83.61</b> | 83.04        | 82.80        |
|            | A+T                           | <b>95.01</b> | <b>84.86</b> | 80.01        | 87.12        | <b>86.75</b> |

Table 3: Performance comparisons of our method for unimodal and multimodal emotion recognition on the IEMOCAP and MSP-Improv datasets. A and T refer to the audio and text modality, respectively. (w/o  $\mathcal{L}_{ua}$ ) means to remove  $\mathcal{L}_{ua}$ .

| Approaches                          | WA(%)        |
|-------------------------------------|--------------|
| bc-LSTM(Poria et al., 2017a)        | 75.60        |
| CATF-LSTM(Poria et al., 2017b)      | 80.10        |
| Zheng.(Lian et al., 2019)           | 78.02        |
| DANN (Lian et al., 2020)            | 82.68        |
| CONSK-GCN(Fu et al., 2021)          | 84.79        |
| Wen. (Wu et al., 2021)              | 83.08        |
| Soumya. (Dutta and Ganapathy, 2022) | 83.80        |
| MER-HAN (Zhang et al., 2023b)       | 73.33        |
| Bubai. (Maji et al., 2023)          | 83.57        |
| Our Method (A)                      | 69.73        |
| Our Method (T)                      | 80.62        |
| <b>Our Method (A+T)</b>             | <b>85.92</b> |

Table 4: Performance comparisons of different methods on the IEMOCAP dataset. A and T refer to the audio and text modality, respectively.

| Approaches              | Metric | A                  | T                  | A+T                |
|-------------------------|--------|--------------------|--------------------|--------------------|
| MCTN(Pham et al., 2019) | F1(%)  | 32.85              | 50.50              | 56.11              |
| MMIN(Zhao et al., 2021) | F1(%)  | 46.47              | 55.73              | 61.88              |
| Bi-LSTM                 | F1(%)  | 44.06 <sup>‡</sup> | 60.04 <sup>‡</sup> | 63.57 <sup>‡</sup> |
| GRU                     | F1(%)  | 43.49 <sup>‡</sup> | 73.92 <sup>‡</sup> | 73.18 <sup>‡</sup> |
| Bi-GRU                  | F1(%)  | 49.08 <sup>‡</sup> | 82.27 <sup>‡</sup> | 82.70 <sup>‡</sup> |
| <b>Our Method</b>       | F1(%)  | <b>55.50</b>       | <b>82.68</b>       | <b>85.26</b>       |

Table 5: Performance comparisons of different methods on the MSP-Improv dataset. <sup>‡</sup> indicates the obtained results of reproducing the corresponding methods. A and T refer to the audio and text modality, respectively.

# Experiments

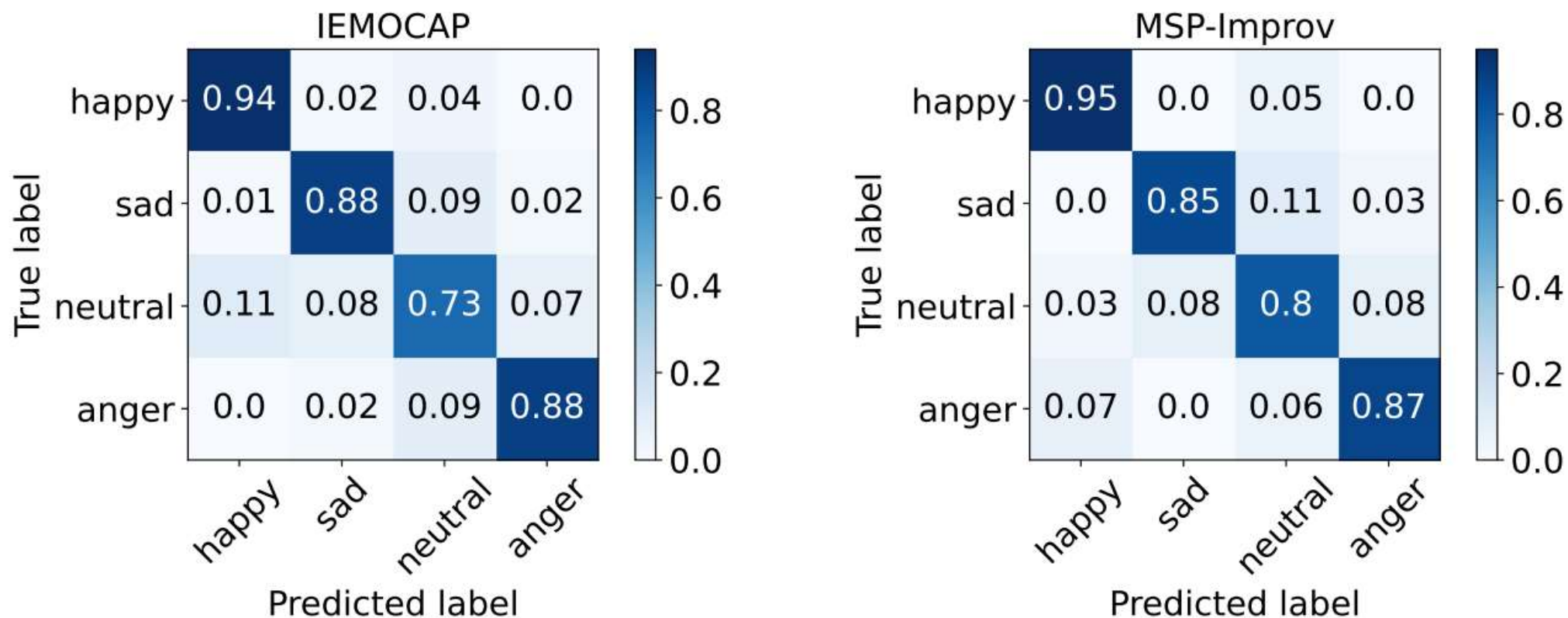


Figure 3: The confusion matrices of recognition results obtained by our method on two datasets: (left) IEMOCAP, (right) MSP-Improv.

# Experiments

| Dataset    | $\mathcal{L}_{ua}$ | $\mathcal{L}_{sp}$ | Cross-modality Attention | WA(%)        | F1(%)        |
|------------|--------------------|--------------------|--------------------------|--------------|--------------|
| IEMOCAP    | ✓                  | ✓                  | ✓                        | <b>85.92</b> | <b>84.48</b> |
|            | ✓                  | ✓                  | ✗                        | 85.23(↓)     | 83.96(↓)     |
|            | ✓                  | ✗                  | ✗                        | 84.79(↓)     | 83.02(↓)     |
|            | ✗                  | ✗                  | ✗                        | 80.85(↓)     | 78.91(↓)     |
| MSP-Improv | ✓                  | ✓                  | ✓                        | <b>86.75</b> | <b>85.26</b> |
|            | ✓                  | ✓                  | ✗                        | 86.13(↓)     | 84.83(↓)     |
|            | ✓                  | ✗                  | ✗                        | 84.64(↓)     | 84.29(↓)     |
|            | ✗                  | ✗                  | ✗                        | 79.36(↓)     | 78.33(↓)     |

Table 6: The effect of key components in our method.

| Dataset    | Methods    | w/o $\mathcal{L}_{ua}$ | w/ $\mathcal{L}_{ua}$ |
|------------|------------|------------------------|-----------------------|
|            |            | WA (%) / F1 (%)        | WA (%) / F1 (%)       |
| IEMOCAP    | Bi-LSTM    | 71.36/69.23            | 73.21/71.28           |
|            | Bi-GRU     | 83.14/82.01            | 84.64/83.63           |
|            | Our Method | 81.84/80.51            | <b>85.92/84.48</b>    |
| MSP-Improv | Bi-LSTM    | 69.54/67.63            | 71.26/69.28           |
|            | Bi-GRU     | 83.35/82.64            | 84.93/84.05           |
|            | Our Method | 82.80/81.34            | <b>86.75/85.26</b>    |

Table 7: The effect of unimodal auxiliary loss ( $\mathcal{L}_{ua}$ ) on the IEMOCAP and MSP-Improv datasets.

# Experiments

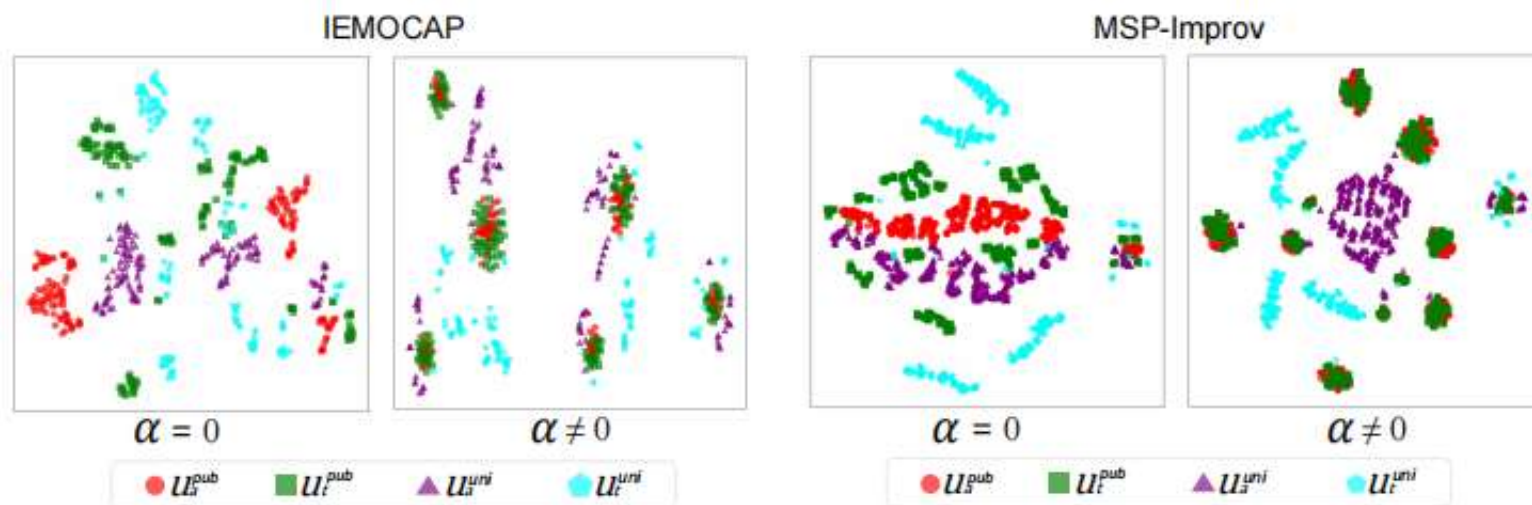


Figure 4: t-SNE visualizations of mapped features: (left) IEMOCAP, (right) MSP-Improv.

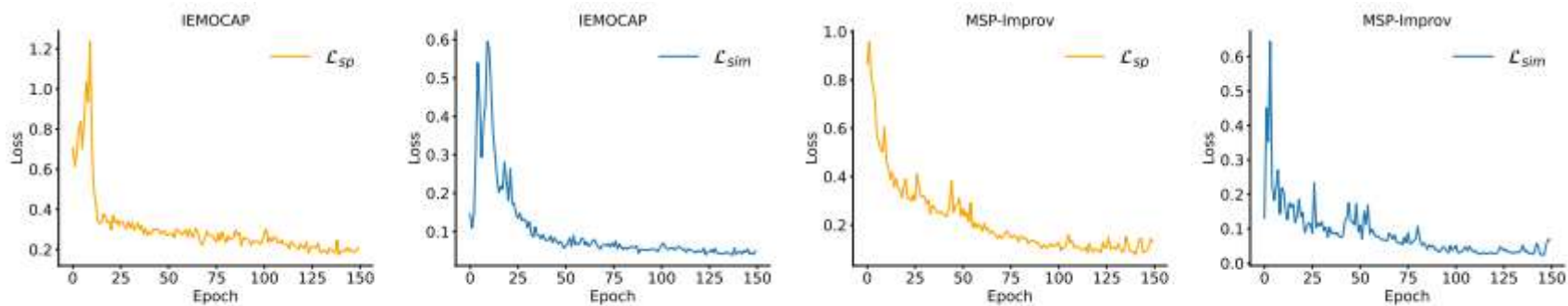


Figure 5: Trends in the regularization loss during training on the IEMOCAP and MSP-Improv datasets.

**Thank you for your  
attention.**