# Two-Stream Temporal Convolutional Network for Dynamic Facial Attractiveness Prediction

**Author**: Nina Weng; Jiahao Wang; Annan Li; Yunhong Wang
**Presenter**: Nina Weng

25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

# Problem Definition

## Dynamic Facial Attractiveness Prediction

Static faces

Dynamic faces

Why meaningful?

- **Psychological and neuroscience**: temporal cues play an important role in the perception of human faces.
- **Industry demand**: increasing popularity of short video apps (tens of thousands of facial performance videos are uploaded per day in Tik Tok).

Highlight of our work

- Propose the dynamic facial attractiveness prediction problem in short videos;
- VFAP Dataset is introduced to facilitate related studies;
- 2S-TCN model is introduced to explore facial appearance and landmark features simultaneously;
- Extensive experiments on VFAP to explore DFAP problem.

# Building Dataset

**Source** — TikTok

**Criterions** — 5 criterions for filtering short videos (detailed in our article)
GOAL: minimize the influence of other factors on facial attractiveness assessment

**Attractiveness Score**

**Automatic scoring** based on interactive behaviors on platform

- Number of likes ↑
- Number of comments ↑
- Number of forwards ↑
- Upload days ↓

$$score = \ln\left(\frac{n_{likes} + n_{comments} + n_{forwards}}{upload\_days}\right)$$

# VFAP Dataset

IRIP laboratory

BASIC STATISTICS OF VFAP DATASET

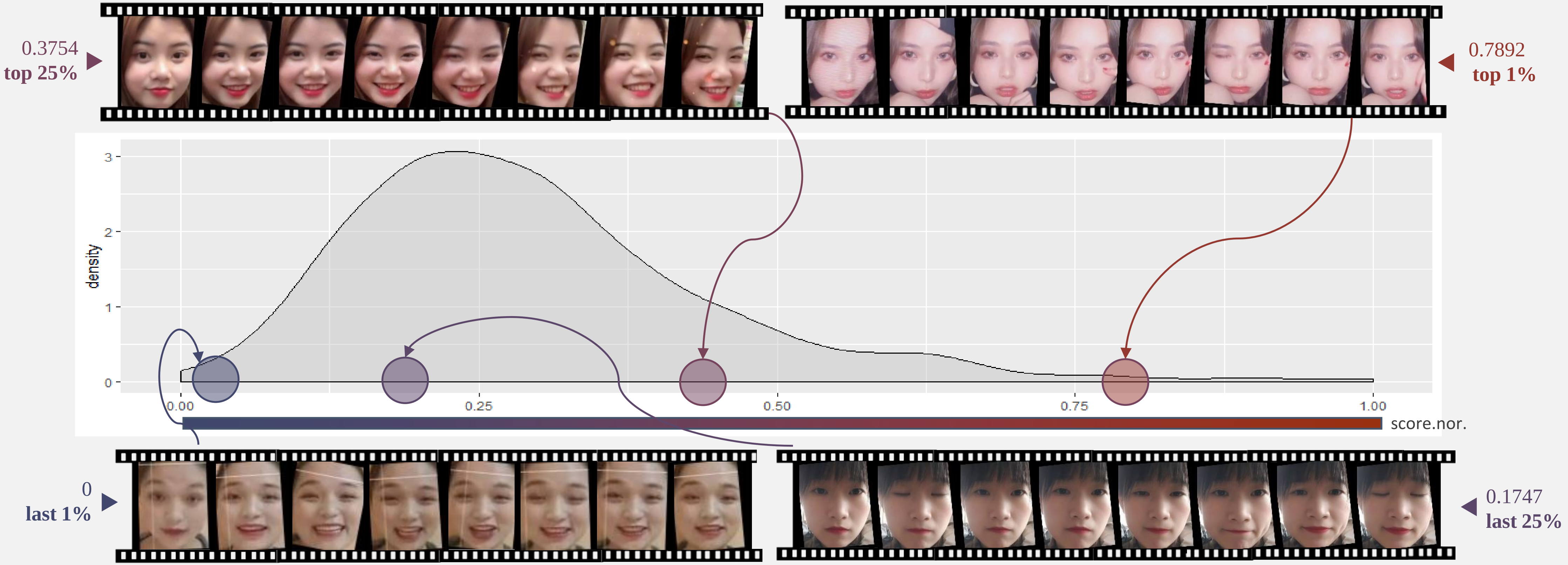
| | S0 | S1 | S2 | S3 | Total |
|---|---|---|---|---|---|
| Video number | 192 | 707 | 365 | 175 | 1,430 |
| Avg. length (second) | 10.79 | 10.99 | 11.99 | 11.35 | 11.26 |
| Total length (minute) | 34.54 | 129.47 | 71.13 | 33.12 | 268.26 |
| Avg. frames | 323.10 | 320.85 | 302.62 | 302.24 | 314.34 |
| Total frames | 62,036 | 22,6841 | 107,734 | 52,892 | 449,503 |

- The VFAP dataset contains 1,430 short videos
- Avg. length for each video is around 11 seconds, and the total length is 268 minutes.
- The dataset is divided into four subsets according to the different TikTok channels.

The introduction video of VFAP dataset is strongly recommended for overall understanding of the dataset and the task.

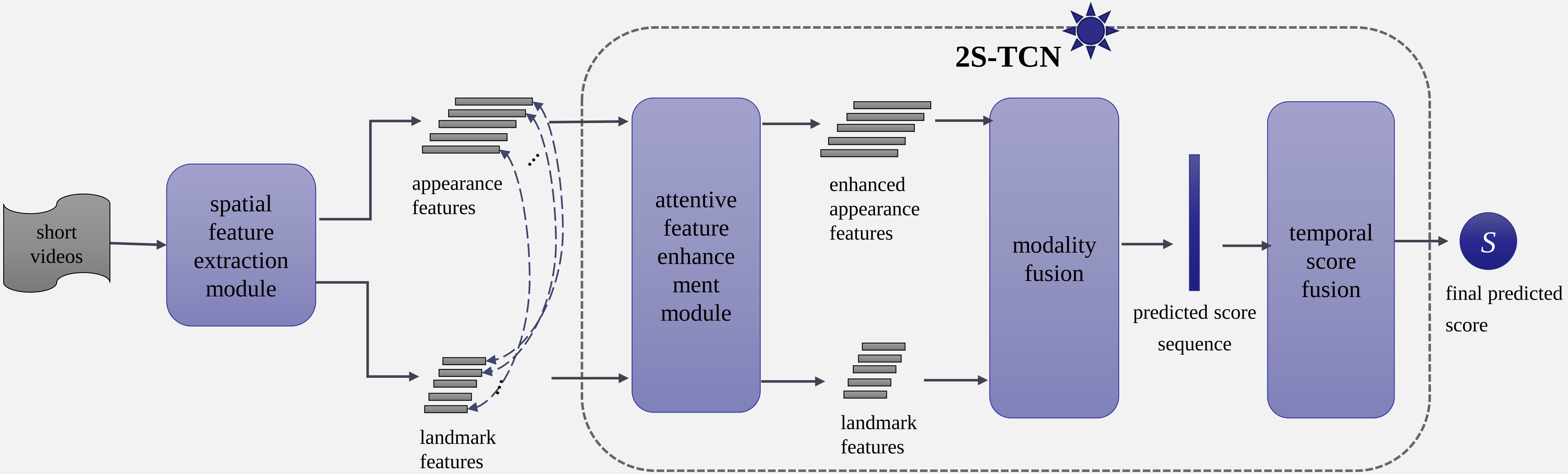
# VFAP Dataset

Distribution of the automatically generated score (with examples)

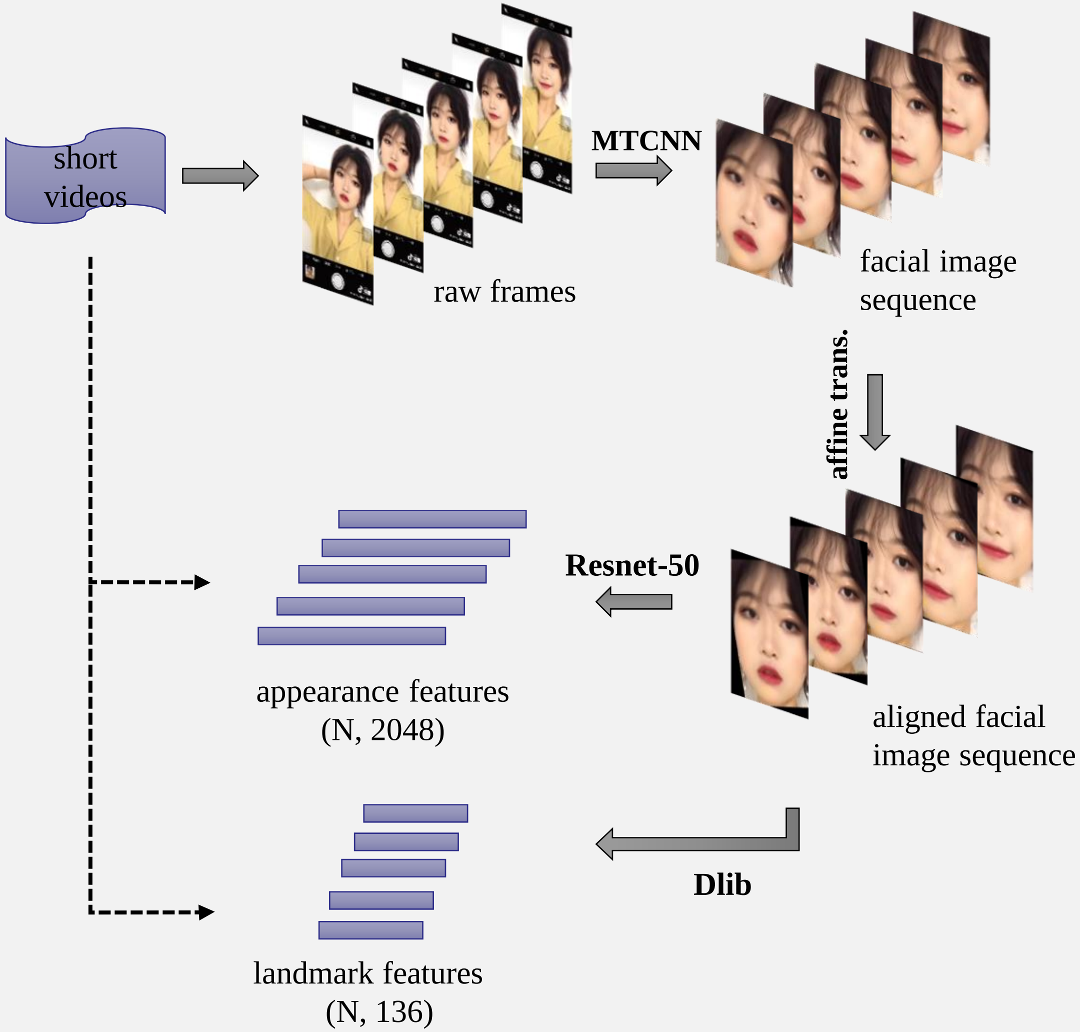


# Framework

Overview of our proposed framework

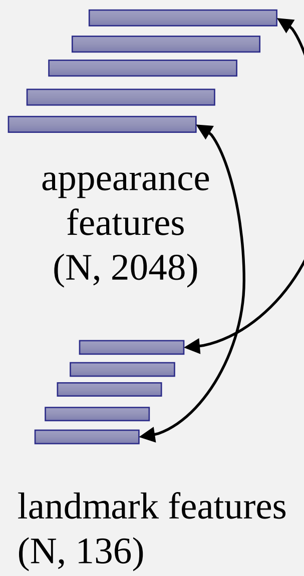


# Spatial Feature Extraction Module

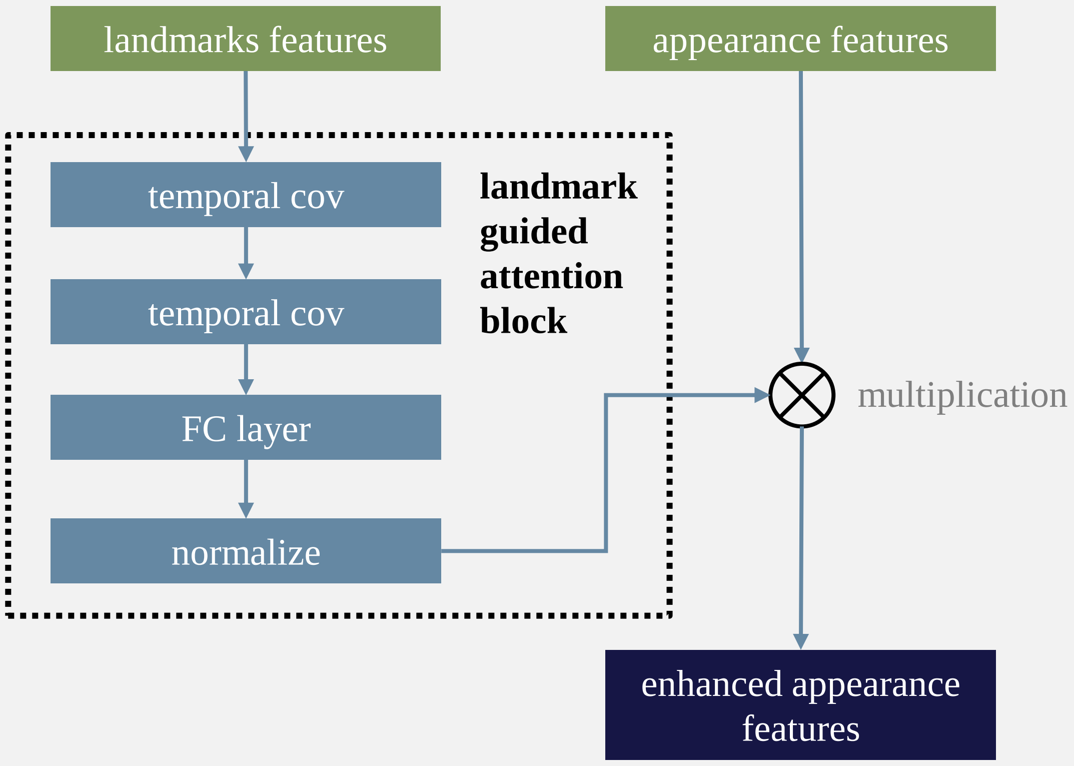



- Goal: generating uniformed representation for each video spatially
- 2 modalities are extracted:
  - (a) facial appearance
  - (b) landmark positions
- State-of-the-art deep models are used for better feature extraction

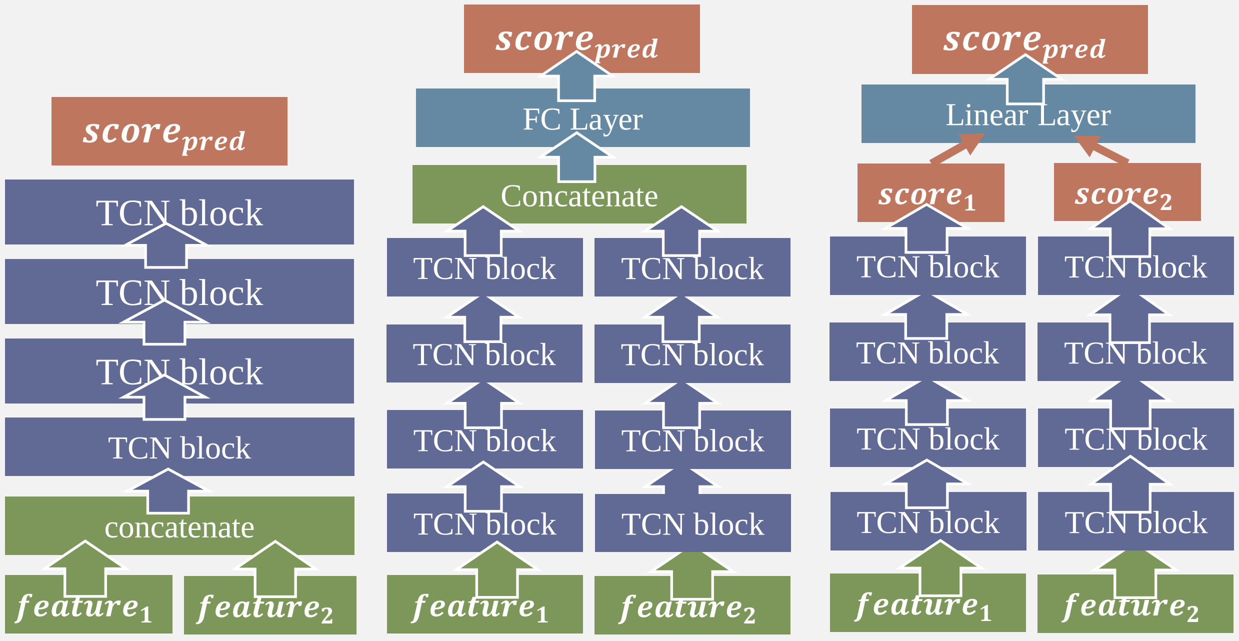
# Two-Stream Temporal Convolutional Network - Attentive feature enhancement




There is a one-to-one correspondence relationship between two extracted features at any time point $t$.

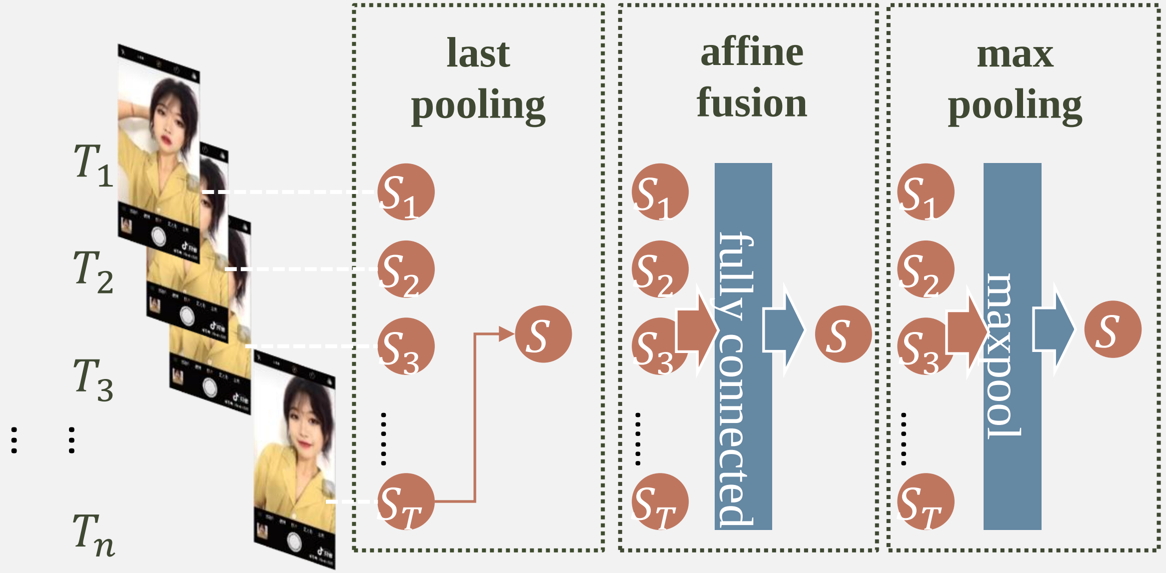


# Two-Stream Temporal Convolutional Network -
## Modality fusion & Temporal score fusion




(a) fusion at data-level (b) fusion at decision-level (c) fusion at score-level

Modality fusion




Temporal score fusion

# Performance Comparison

**Evaluation Metrics**
Spearman's rank correlation (SRC) coefficient

$$\rho = \frac{\sum_i (S_i-\bar{S})(R_i-\bar{R})}{\sqrt{\sum_i (S_i-\bar{S})^2 \sum_i (R_i-\bar{R})^2}} = 1 - \frac{6\sum_i {d_i}^2}{n(n^2-1)}$$

**Loss function:**
the weighted summation of mean squared error (MSE) $L_m$ and margin ranking loss $L_r$.

$$Lm = \frac{1}{n}\sum_{i=1}^{n}(p_i-g_i)^2$$
$$L_r = \sum_{i=1}^{n}\sum_{j=1,j>i}^{n} \max((p_i-p_j)\times sign(g_j-g_i)+\delta,0)$$
$$L = L_m + \alpha L_r$$

TABLE II
COMPARISON WITH STATIC FAP METHODS ON VFAP.

| Method | S0 | S1 | S2 | S3 | All |
|---|---|---|---|---|---|
| AlexNet | 0.08920 | 0.01559 | 0.03243 | 0.13180 | 0.01388 |
| Resnet-18 | 0.11078 | 0.00787 | 0.03016 | 0.12721 | -0.00240 |
| ResNeXt-50 | 0.12934 | 0.06878 | 0.02485 | 0.12898 | -0.00388 |
| 2S-TCN | **0.38621** | **0.26273** | **0.32138** | **0.38699** | **0.18965** |

- Static methods do not predict well on the dynamic facial data
- -> the importance of temporal modeling for dynamic facial data

IRIP laboratory

# Ablation study

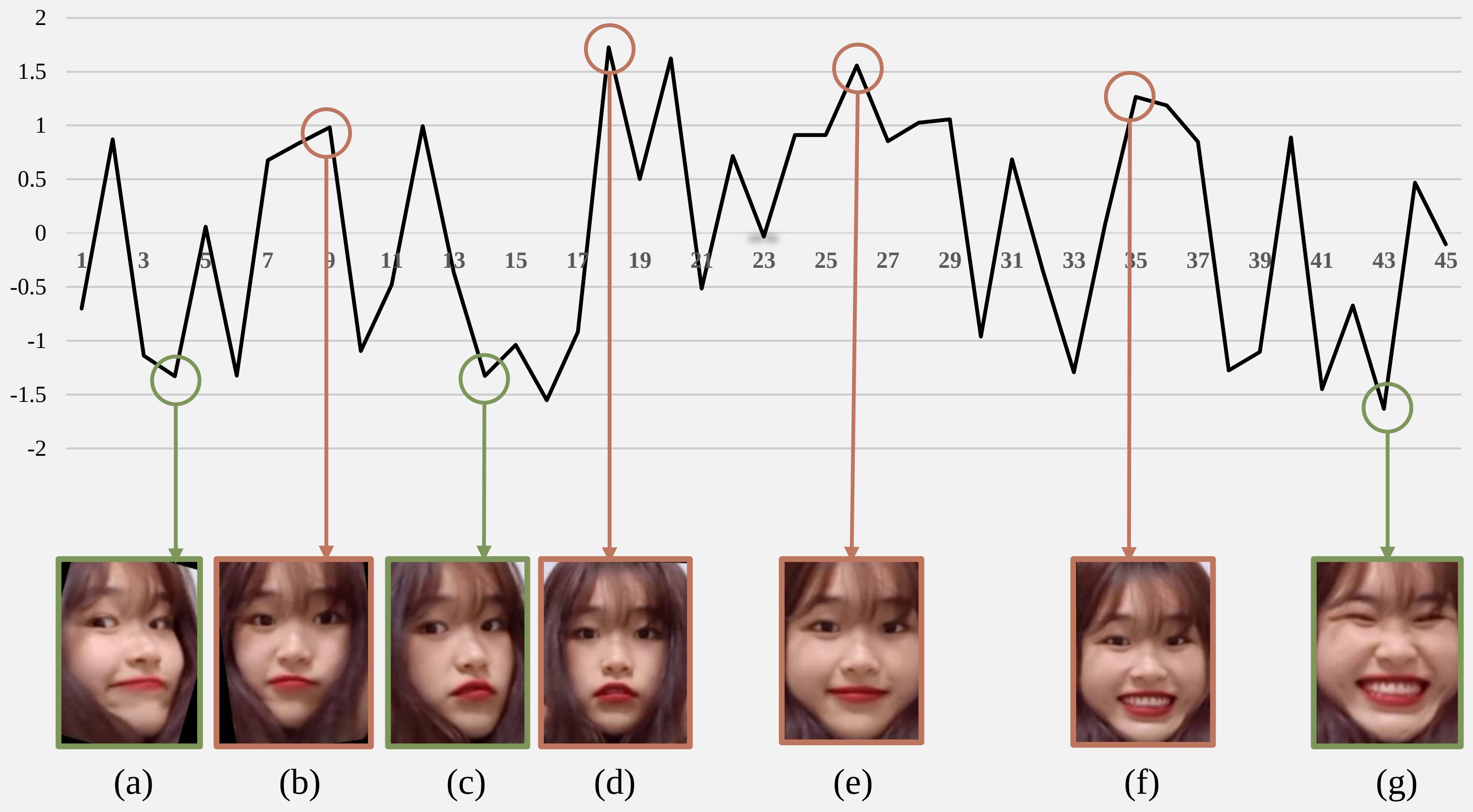
The proposed 2S-TCN structure which adopts **attentive feature enhancement**, **decision-level fusion** and **max pooling** achieves the best outcome.

TABLE III
RESULTS OF ABLATION STUDIES ON 2S-TCN MODEL.

| modality | | attention | modality fusion | | | score fusion | | | subsets | | | | all |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| appearance | landmarks | | feature-level | decision-level | score-level | last | affine | max | S0 | S1 | S2 | S3 | |
| ✓ | | | | | | | | ✓ | 0.2784 | 0.2162 | 0.2359 | 0.2769 | 0.1606 |
| | ✓ | | | | | | | ✓ | 0.3064 | 0.1995 | 0.2568 | 0.3026 | 0.1602 |
| ✓ | ✓ | | ✓ | | | | | ✓ | 0.3122 | 0.2235 | 0.2778 | 0.3170 | 0.1692 |
| ✓ | ✓ | | | ✓ | | | | ✓ | 0.3380 | 0.2209 | 0.3018 | 0.3077 | 0.1820 |
| ✓ | ✓ | | | | ✓ | | | ✓ | 0.3331 | 0.2133 | 0.2830 | 0.3255 | 0.1696 |
| ✓ | ✓ | ✓ | ✓ | | | | | ✓ | 0.3847 | 0.2478 | 0.3065 | 0.3269 | 0.1693 |
| ✓ | ✓ | ✓ | | ✓ | | | | ✓ | **0.3848** | **0.2498** | **0.3071** | 0.3272 | **0.1908** |
| ✓ | ✓ | ✓ | | | ✓ | | | ✓ | 0.3548 | 0.2309 | 0.2981 | **0.3360** | 0.1712 |
| ✓ | ✓ | ✓ | | ✓ | | | ✓ | | 0.3294 | 0.2225 | 0.3066 | 0.3236 | 0.1699 |
| ✓ | ✓ | ✓ | | ✓ | | ✓ | | | 0.3169 | 0.2163 | 0.2720 | 0.3209 | 0.1671 |

# Qualitative analysis

## - Attentions generated from attentive feature enhancement module




**Attention reducing factors:**

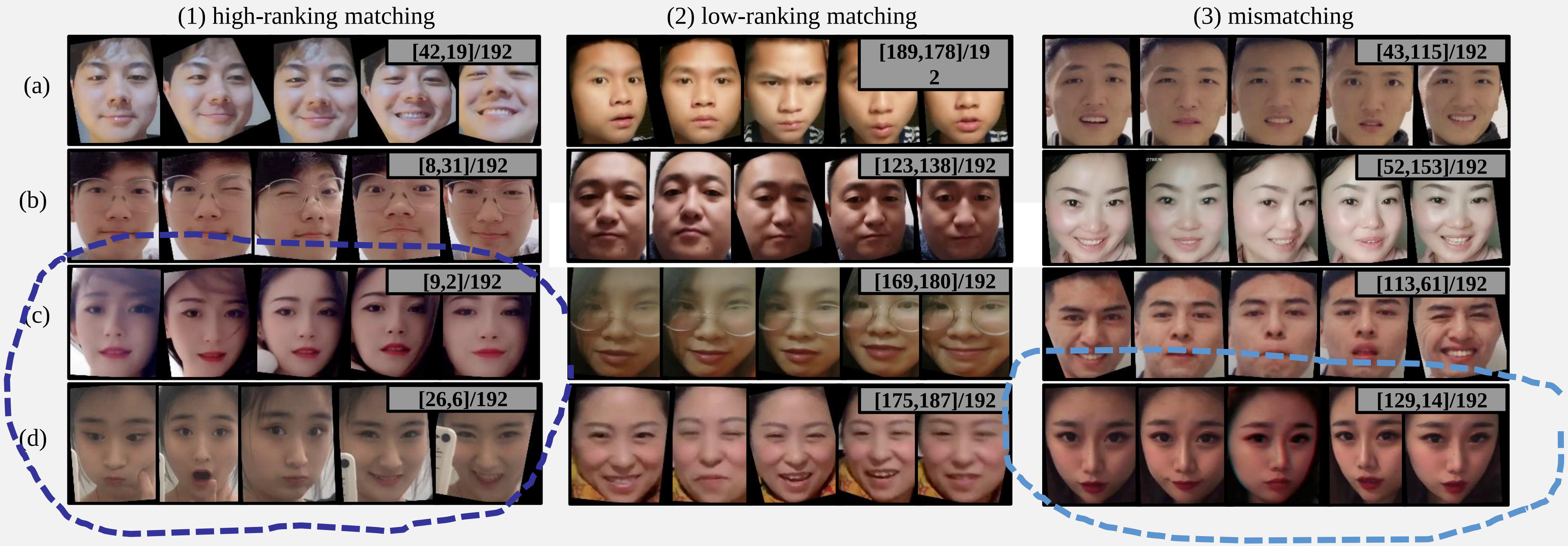
- not frontal face (a,c)
- too exaggerated or even distorted facial expressions (g)

**Attention increasing factors:**

- Positive facial expressions (e,f)

# Qualitative analysis

## - Matching and mismatching examples




- two different kinds of dynamic facial attractiveness, i.e. beauty in facial appearance (example (1; c)) and interestingness in facial expressions (example (1; d)).
- the low-ranking faces lack attractiveness in both facial appearance and expressions.

**Problems and future improvements:**

- the deviations in attractiveness scores → a better formulation of the attractiveness score
- bias in the gender distribution → intentionally introducing more representative male videos into dataset

IRIP laboratory

# Two-Stream Temporal Convolutional Network for Dynamic Facial Attractiveness Prediction

# Thanks for watching!

Contacts: Nina Weng {wengnn@buaa.edu.cn}

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021