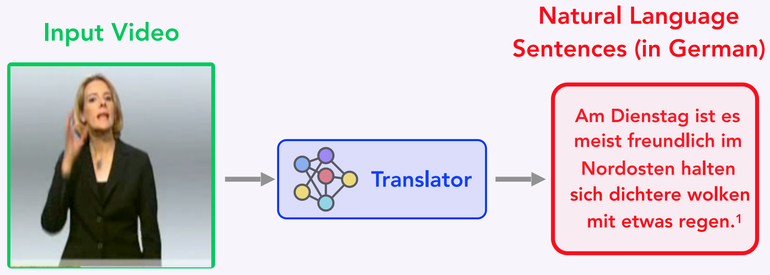
TSPNet: Hierarchical Feature Learning via Temporal Semantic Pyramid for Sign Language Translation

Dongxu Li*, Chenchen Xu*, Xin Yu, Kaihao Zhang, Ben Swift, Hanna Suominen, Hongdong Li

NEURAL INFORMATION PROCESSING SYSTEMS

Neural Sign Language Translation



1. Translation in English: "On Tuesday it is mostly friendly in the Northeast there are thicker clouds with some rain".

Research Background

- **Deaf and hard-of-hearing population** worldwide (466M) have difficulties in public involvement and career development.
- Sign language is the **primary communication channel** in the deaf communities around the world.
- **Neural sign language translation (NSMT)** is a challenging multi-modality sequence-to-sequence task.
- **Applicable NSMT scenarios** include hospital and restaurant services, career consultation and social welfare.

Motivations and Contributions

Motivations

- Current approaches extract sign features in a frame-wise fashion, thus are disadvantageous in capturing temporal dependencies.
- Obtaining accurate sign gesture segmentation is difficult without laborious per-frame annotations.

Observations on sign video semantics

- **Coherence** - temporally neighboring sign video segments are semantically consistent.
- **Context-dependency** - non-local video context influences the interpretation of individual sign gestures.

Contributions

- **Multi-scale segment representation** - to better model temporal information of sign gestures.
- **Local feature learning** - to enforce semantic consistency.
- **Non-local feature learning** - contextual semantic disambiguation.

TSPNet - Temporal Semantic Pyramid Network

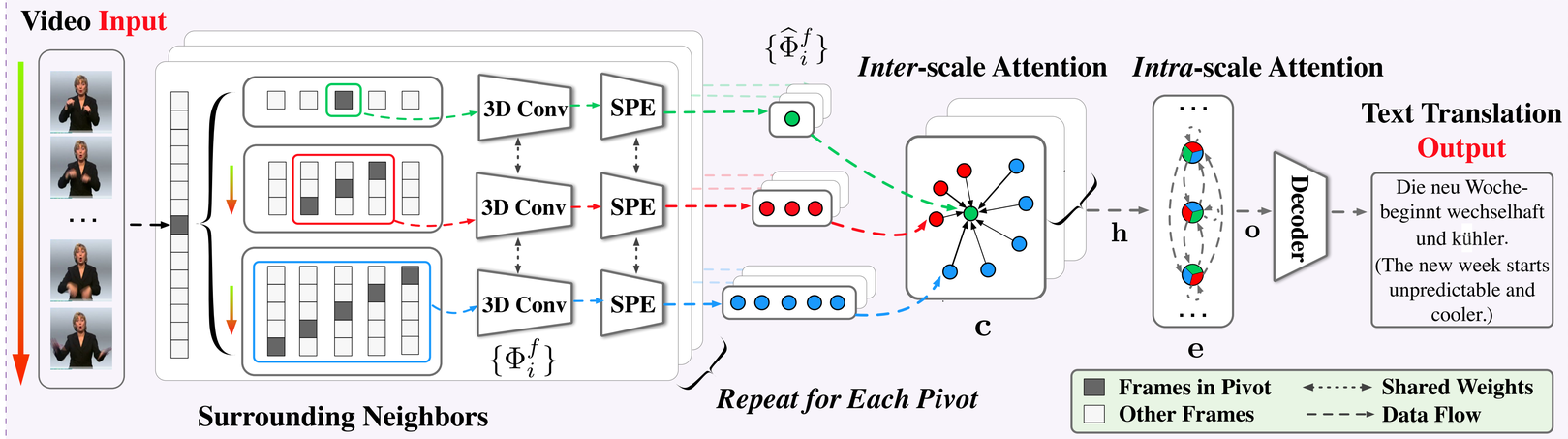


Fig. 1. Overview of the TSPNet workflow, which generates natural language translations directly from sign language videos.

Multi-scale Segment Representation and Surrounding Neighborhood

- Given a sign language video, we first generate **windowing segments of different temporal granularities** to account for gesture of different lengths;
- **Shared position encoding** to inform sequence information memory-efficiently;
- Take each segment in the smallest scale as a **pivot segment**, construct its **neighborhood segments** on larger scales that superset its frames.
- **Structured attention operations** to enforce local semantic consistency and resolve non-local semantic ambiguity.

Formulations

Pivot Segment — Surrounding Neighbourhood

Local neighborhood - the set of multi-scale segments that superset the pivot. $\mathcal{Z}_k^* = [\mathbf{W}_c z_0^*, \mathbf{W}_c z_1^*, ..., \mathbf{W}_c z_{Q-1}^*]^T$ where $z_{j>0}^* \in \mathcal{N}_{\widehat{\phi}_{0,k}^f}^*$, $Q = |\mathcal{N}_{\widehat{\phi}_{0,k}^f}^*|$

Joint local and non-local feature learning - structured attention over neighborhood. $c_k^* = \mathcal{G}_{\text{attn}}(\mathcal{Z}_k^*, \mathcal{Z}_k^*, \mathcal{Z}_k^*)$.

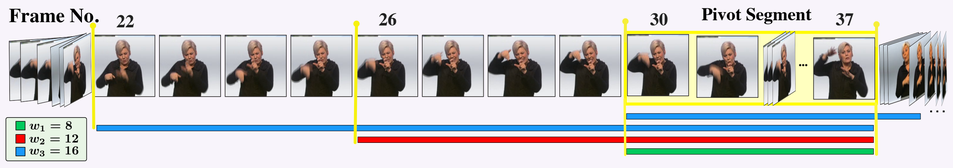


Fig 2. A surrounding neighborhood consists of features of a pivot segment (30th-37th frames) and neighbouring segments from multiple scales. Here we show four segments with the highest inter-scale attention scores. From the 26th to 37th frame, the segment corresponds to the GSL sign word **süden** (south).

Qualitative Results

Methods	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Conv2d-RNN	29.70	27.10	15.61	10.82	8.35
+ Luong Attn.	30.70	29.86	17.52	11.96	9.00
+ Bahdanau Attn.	31.80	32.24	19.03	12.83	9.58
TSPNet-Sequential	**34.77**	**35.56**	**22.80**	**16.60**	**12.97**
TSPNet-Joint	**34.96**	**36.10**	**23.12**	**16.88**	**13.41**

Quantitative Results

Ground Truth:	der wind weht meist schwach aus unterschiedlichen richtungen. (mostly windy, blowing in weakly from various directions.)
Conv2d-RNN:	**der wind weht schwach** bis mäßig . (**windy, blows weak** to moderate .)
Ours:	**der wind weht meist schwach aus unterschiedlichen richtungen.** (**mostly windy, blowing in weakly from various directions.**)
Ground Truth:	im süden und südwesten gebietsweise regen sonst recht freundlich. (in the south and southwest locally rain otherwise quite friendly.)
Conv2d-RNN:	von der südhälfte beginntes vielerorts. (from the southpart it starts in many places.)
Ours:	**im süden** gibt es heute nacht noch einzelne **schauer**. **In the south** there are still **some showers tonight**.

Correct 1-grams. Correct semantics.