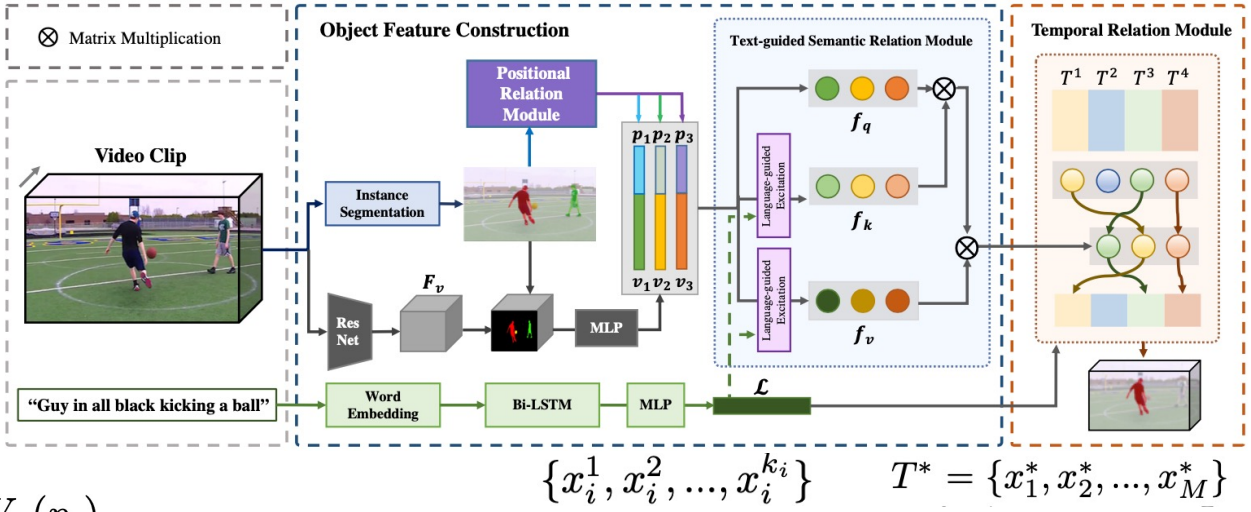
ClawCraneNet [Liang+ (Univ. of Technology Sydney), 21]

- Text-based Video Segmentation taskに焦点
  - 各フレームで全オブジェクトのマスクから言語情報で参照されるものを選択
  - Referring Youtube - VOS challengeで1位
- Object features $v^j = \mathrm{MLP}(\mathrm{Max}(F_v \odot o^j))$
  - jは候補物体(N_v個)、Fvは画像全体の特徴量
- Positional Relation Module

$p_i = (x^i_{min}, y^i_{min}, x^i_{max}, y^i_{max}, x^i_c, y^i_c, w_i, h_i, r^x_i, r^y_i) \quad \mathcal{V}_i = v_i + W_p(p_i)$

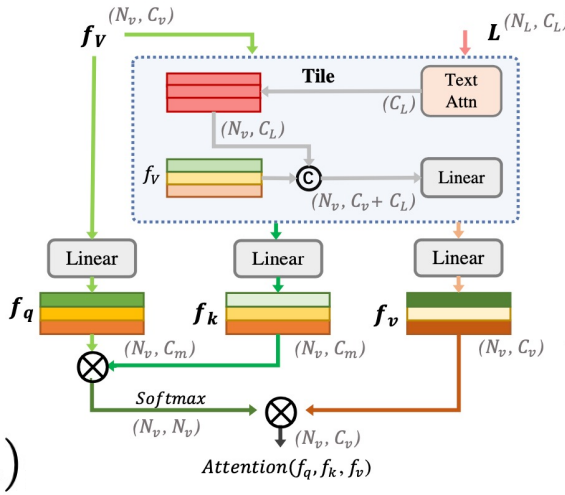
  - o^iの最小外接boxの左上、右下、中心の座標、幅、高さ
  - r^x_iとr^y_iは、x, y軸の相対位置のインデックス
- Temporal Relation Module $\mathcal{S}_s(x_i, x_j) = \mathcal{S}_c(\mathcal{V}^r_i, \mathcal{V}^r_j) + \alpha * U(x_i, x_j)$
  - 隣接フレームからの2オブジェクトx_i, x_j
  - Sc:コサイン類似度、U:IoUの重なり
  - 値γよりも大きく、β回のマッチングで更新されない場合に適応
- 言語特徴量Lとのコサイン類似度が最大のV^r_iを採用
  - 損失関数
    - gt:ground truth

$$s_i = \frac{\exp(\mathcal{V}^{r\,\mathrm{T}}_i \mathcal{L}/\tau)}{\sum_{j=1}^{N_v} \exp(\mathcal{V}^{r\,\mathrm{T}}_j \mathcal{L}/\tau)} \quad \mathrm{loss} = -\log(s_{gt})$$



$\{x^1_i, x^2_i, ..., x^{k_i}_i\}$ $T^* = \{x^*_1, x^*_2, ..., x^*_M\}$

Text-guided Semantic Relation Module

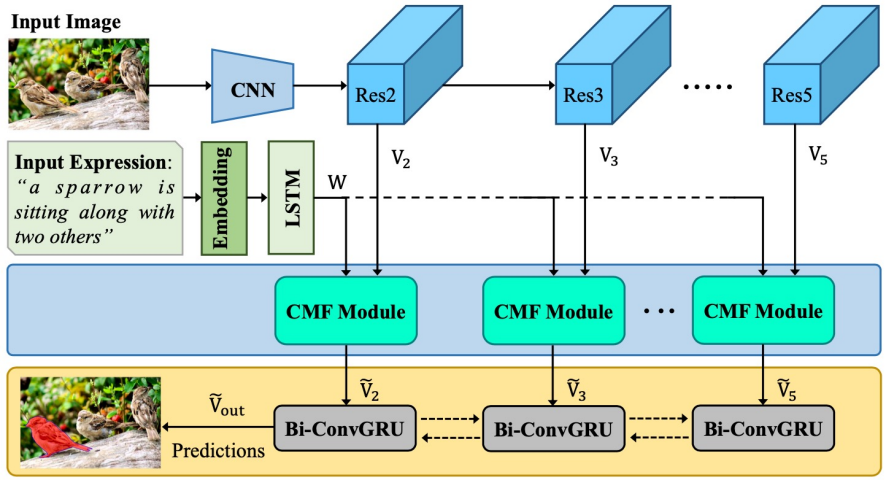


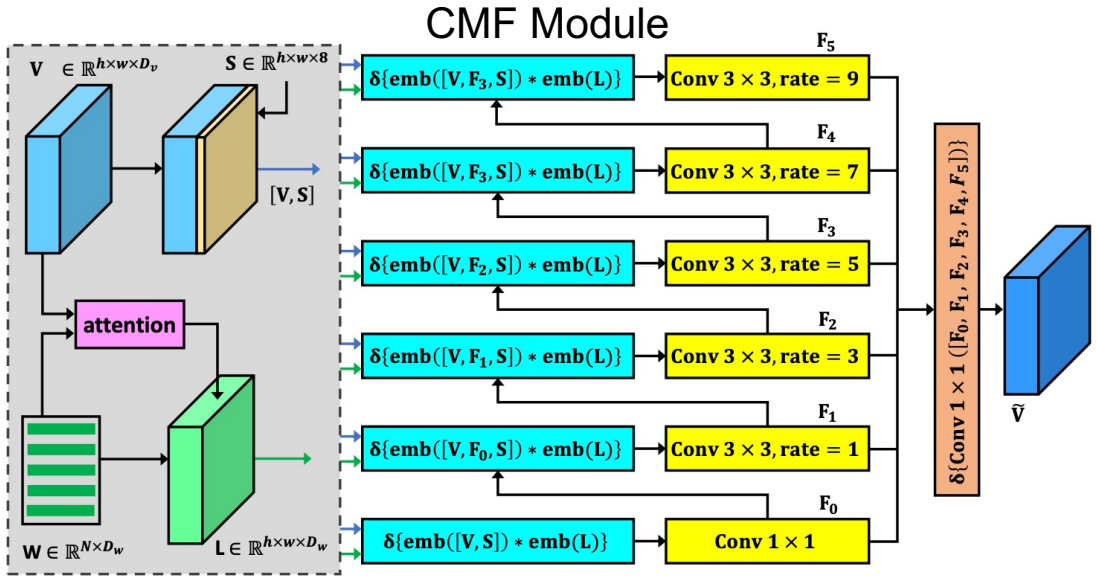
$$f_o = W_o(\mathrm{Concat}(f_V, f_t))$$
$$f_V = (\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_{N_v})$$
$$f_q = W_q f_V$$
$$f_k = W_k f_o$$
$$f_v = W_v f_o$$
$$F_o = f_V + \mathrm{softmax}(\frac{f_q f_k^T}{\sqrt{d_k}}) f_v$$
$$F_o = (\mathcal{V}^r_1, \mathcal{V}^r_2, ..., \mathcal{V}^r_{N_v})$$

Liang, C., Wu, Y., Luo, Y., & Yang, Y. (2021). ClawCraneNet: Leveraging Object-level Relation for Text-based Video Segmentation. arXiv preprint arXiv:2103.10702.

CMF [Yang+ (Beijing Univ.), ICIP21]

- Referring Image SegmentationにおいてG-Ref、UNC、UNC+でSoTA
- Attention Matrix $A_{i,t} = softmax[(W_v v_i)^T (W_h h_t)], \quad l_i = \sum_{t=1}^{T} A_{i,t} h_t$
  - iは視覚特徴のindex、tはt番目の単語の言語特徴
- Cascaded Multi-modal Fusion(CMF)
  - $emb(\cdot)$: 1x1 conv 、δ : ReLU、$*$: アダマール積
- bi-directionally convolutional GRU $\tilde{V}_{out} = ReLU(W_p^{\overrightarrow{H}} \overrightarrow{H}_l + W_p^{\overleftarrow{H}} \overleftarrow{H}_l + b)$
  - マルチモーダルな特徴を
    ボトムアップとトップダウンの2つの方法で統合

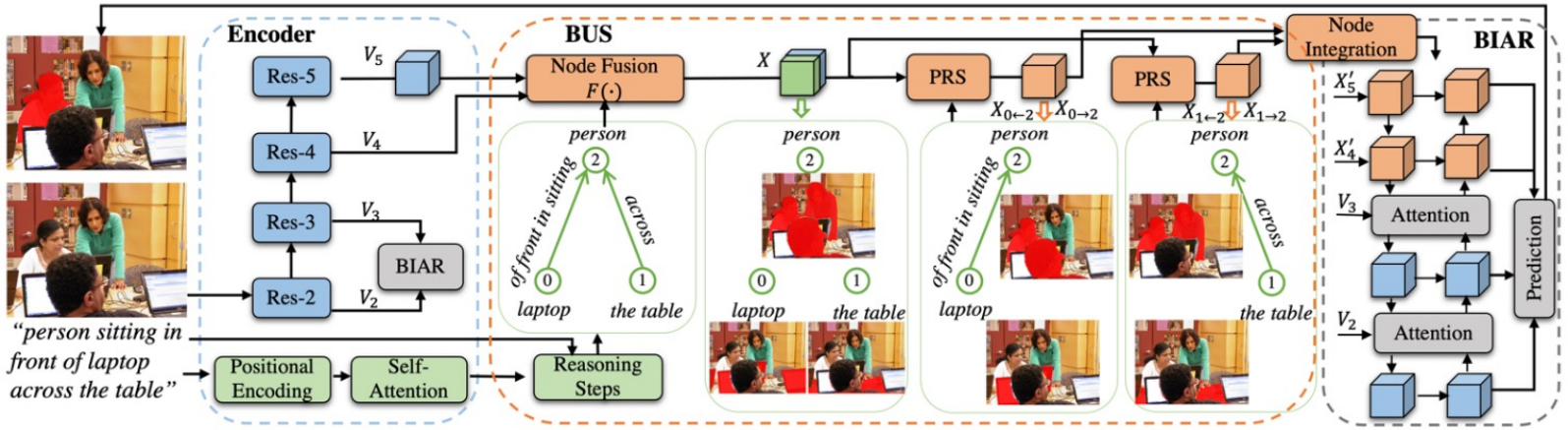




Yang, J., Huang, Y., Ma, Z., & Wang, L. (2021). CMF: Cascaded Multi-model Fusion for Referring Image Segmentation. arXiv preprint arXiv:2106.08617.

BUSNet [Yang+ (ShanghaiTech Univ.), CVPR21]

- 推論の途中経過が可視化できるRISモデル
  - UNC、UNC+、G-RefでSoTA
- Bottom-Up Shift (BUS)
  - グラフを生成
  - ノード o_n / 有向エッジ e_k
    - 名詞句 / 前置詞、動詞
  - マルチモーダル特徴X_nを生成、X'_nにアップデート
    - V:画像特徴、P:位置、$\bar{h}_n$:o_nの言語特徴、$\mathcal{E}_n$: o_nに繋がるエッジ
    - $*$は要素ごとの乗算、;はConcatenate
- Pairwise Relation Shift (PRS)
  - $\odot$はピクセルごとの乗算、γはtanh
- Bidirectional Attentive Refinement (BIAR)
  - $\{V2, V3, V4, X4', X5'\}$を改良し、$\{G1, G2, G3, G4, G5\}$を得る
  - アップサンプリングして合計し予測



$$\boldsymbol{X}_n = Conv_v([\boldsymbol{V};\boldsymbol{P}]) * Tile(\boldsymbol{W}_{\bar{h}}\bar{\boldsymbol{h}}_n) \quad \boldsymbol{X}_n = F(\boldsymbol{V},\boldsymbol{P},o_n)$$

$$\boldsymbol{X}_{n\leftarrow m}, \boldsymbol{X}_{n\rightarrow m} = PRS^{(3)}(\boldsymbol{X}_n, e_k^{(r)}, \boldsymbol{X}'_m),$$

$$\boldsymbol{X}'_n = \frac{\sum_{o_m \in e_k^{(o)} \& e_k \in \mathcal{E}_n} \boldsymbol{X}_{n\leftarrow m} + \boldsymbol{X}_n}{|\mathcal{E}_n| + 1}$$

$$\boldsymbol{A}_{s\leftarrow o} = \gamma(Conv_r^{-1}(\boldsymbol{X}_o)), \boldsymbol{X}_{s\leftarrow o} = F(\boldsymbol{A}_{s\leftarrow o} \odot \boldsymbol{V}, \boldsymbol{P}, e^{(s)}),$$

$$\boldsymbol{A}_{s\rightarrow o} = \gamma(Conv_r(\boldsymbol{X}_s)), \boldsymbol{X}_{s\rightarrow o} = F(\boldsymbol{A}_{s\rightarrow o} \odot \boldsymbol{V}, \boldsymbol{P}, e^{(o)}),$$

$$\boldsymbol{A}_i^{td} = \sigma(Conv_c(Conv_a(\boldsymbol{G}_i) + Conv_b(Up(\boldsymbol{G}_{i+1}^{td}))))$$

$$\boldsymbol{G}_i^{td} = \begin{cases} Conv_i(\boldsymbol{G}_i), & \text{if } i \in \{5\} \\ Conv_i(\boldsymbol{G}_i + Up(\boldsymbol{G}_{i+1}^{td})), & \text{if } i \in \{4\} \\ Conv_i(\boldsymbol{A}_i^{td} \odot \boldsymbol{G}_i + Up(\boldsymbol{G}_{i+1}^{td})), & \text{if } i \in \{1,2,3\} \end{cases}$$

Yang, S., Xia, M., Li, G., Zhou, H. Y., & Yu, Y. (2021). Bottom-Up Shift and Reasoning for Referring Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11266-11275).