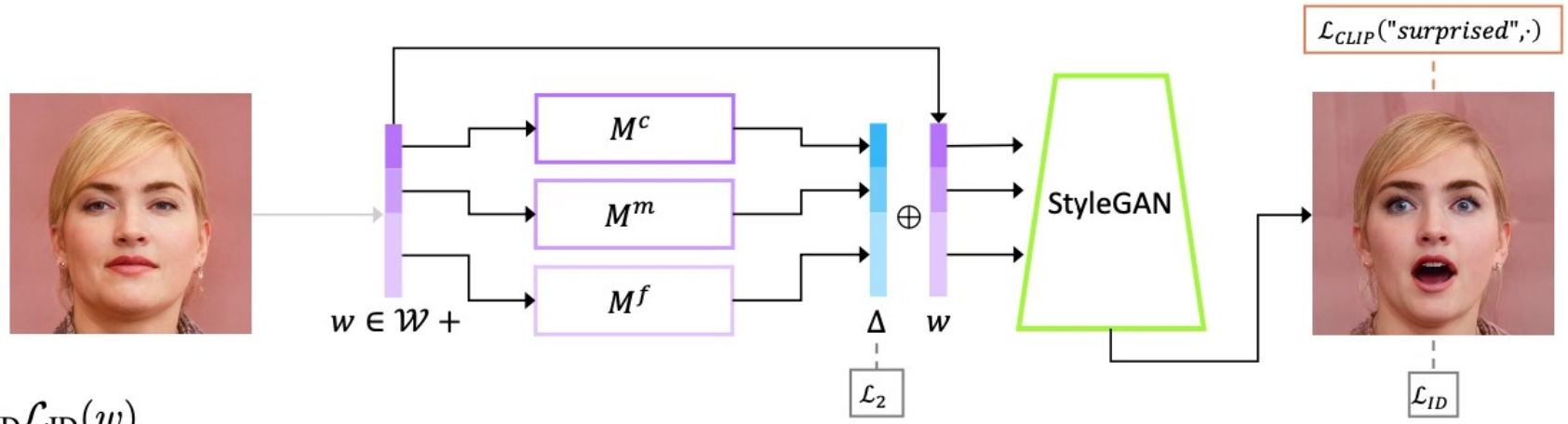
StyleCLIP [Patashnik+ (Adobe Research), CVPR21]

- StyleGANとCLIPを組み合わせた、テキストによる画像操作を行うモデル
- Latent Optimization : $\arg\min D_{\text{CLIP}}(G(w), t) + \lambda_{\text{L2}} \|w - w_s\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w)$
  - D_{CLIP} : CLIP埋め込み空間における、生成画像とテキストのコサイン距離
  - L2 : 元のwとの距離、入力画像との乖離を防ぐ
  - L_{ID} : 入力・生成画像に対して生成した埋め込みのコサイン類似度　$\mathcal{L}_{\text{ID}}(w) = 1 - \langle R(G(w_s)), R(G(w)) \rangle$
    - RはArcFace、人物の乖離を制御
- Latent Mapper
  - wは3つのmapperに入力
    - coarse, medium, fine

$$w = (w_c, w_m, w_f)$$

$$M_t(w) = (M_t^c(w_c), M_t^m(w_m), M_t^f(w_f))$$



  - 損失関数

$$\mathcal{L}(w) = \mathcal{L}_{\text{CLIP}}(w) + \lambda_{L2} \|M_t(w)\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w)$$

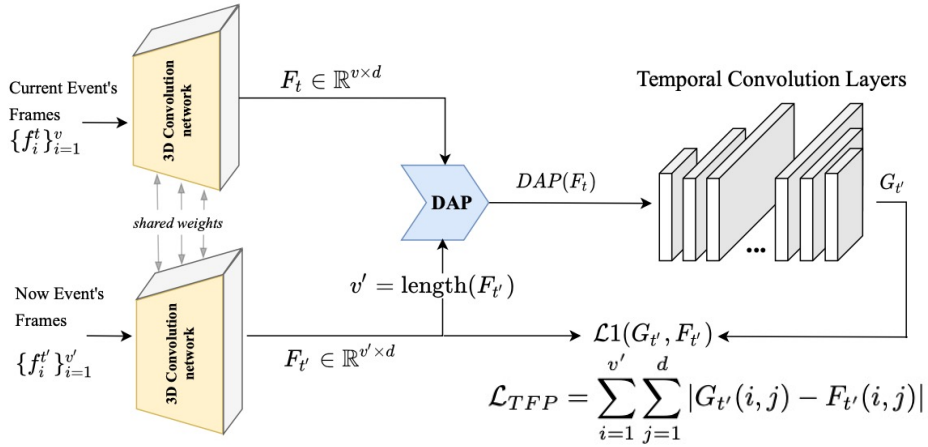
$$\mathcal{L}_{\text{CLIP}}(w) = D_{\text{CLIP}}(G(w + M_t(w)), t)$$

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. arXiv preprint arXiv:2103.17249.

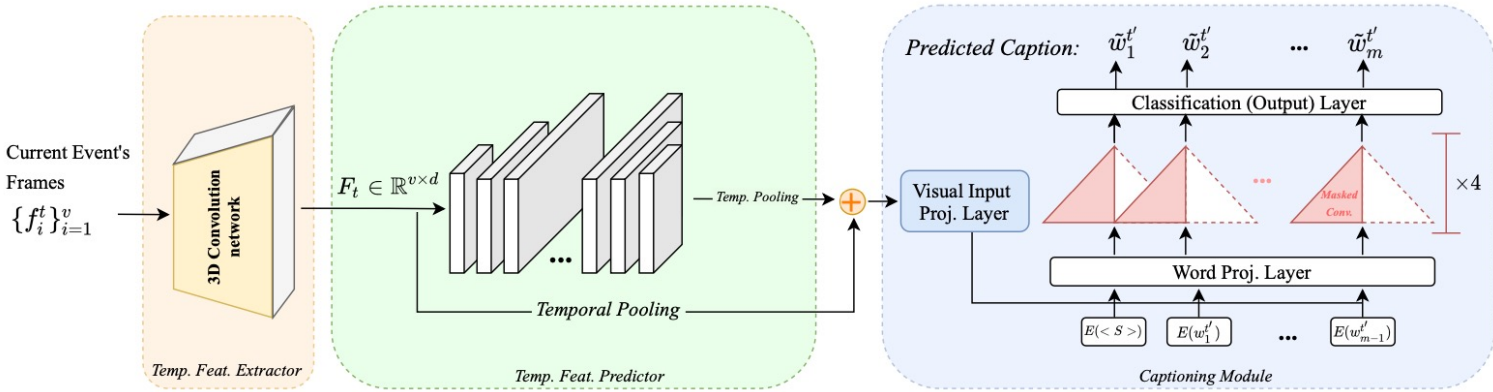
Video Captioning of Future Frames [Hosseinzadeh+ (University of Manitoba), WACV21]

- 未来の出来事にキャプションを付けるタスク
- 次のイベントの畳み込み特徴量を予測し、その特徴量に基づいてキャプションを生成
- Temporal Feature Predictor
  - t番目を元にt'番目のイベントの画像特徴量を予測
  - $G_{t'} = Conv(DAP(F_t)), \text{where } G \in \mathbb{R}^{v' \times d}$
    - v:フレーム数　d:次元数

    $G_{t'}^{final} = \lambda \cdot \mathcal{AVG}(F_t) \oplus (1-\lambda) \cdot \mathcal{AVG}(G_{t'})$
    - AVG:時間次元でのAvg Pooling　⊕:要素ごとの和



- Captioning Module
  - m-1番目までの単語とGt'から m番目の単語を予測
    - classification layerの上に softmaxを使用
  - Lossはcross-entropy



Hosseinzadeh, M., & Wang, Y. (2021). Video Captioning of Future Frames. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 980-989).

HAMLET [Islam+ (Univ. of Virginia), IROS20]

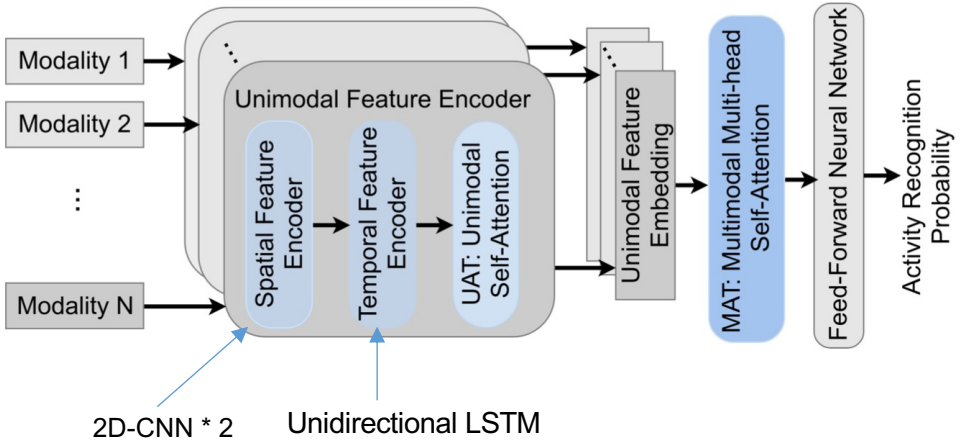
- human activity recognition (HAR)において、テストした全データセットでベースラインを上回る
  - ユニモーダルなデータから特徴量を抽出し、それらを分離・融合するマルチモーダルなAttentionメカニズム
    - HAMLET：Hierarchical Multimodal Self-attention based HAR
- Input：X^m ($B \times S^m \times E^m$)
  - B：バッチサイズ、S^m：モダリティ m のセグメント数、E^m：特徴次元$(channel(C^m) \times height(H^m) \times width(W^m))$
- Unimodal Self-Attention (UAT)
  - H^m：Temporal Feature Encoderの出力

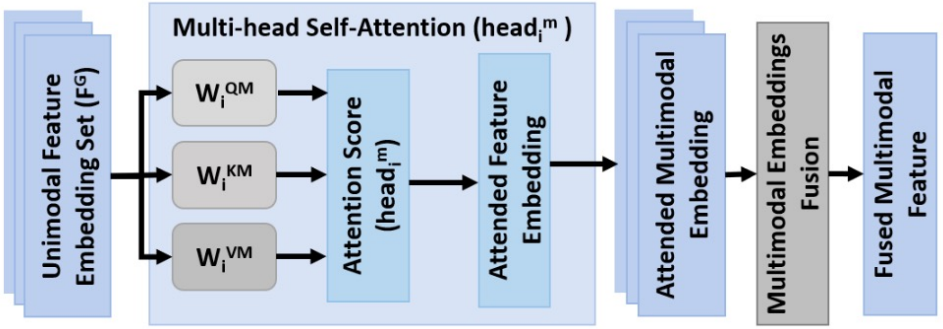
$$Q_i^m = H^m W_i^{Q,m} \quad K_i^m = H^m W_i^{K,m} \quad V_i^m = H^m W_i^{V,m}$$

$$head_i^m = Attn(Q_i^m, K_i^m, V_i^m) \quad F_m^a = [head_1^m; ...; head_h^m] W^{O,m}$$

$$F_m = \sum_{s \in S^m} F_{m,s}^a$$

- Multimodal Attention based Feature Fusion (MAT)
  - F^{Gu}にmulti-head self-attention、出力はF^{Ga}
    - $F^{G^u} = (F_1, F_2, ..., F_M)$ Mは全モダリティ数、順不同
- F^{Ga}から2つの方法でF^Gを算出（CONCATの方が高性能）
  - MAT-SUM $F^G = \sum_{m=1}^{M} F_m^{G^a}$
  - MAT-CONCAT $F^G = [F_1^{G^a}; F_2^{G^a}; ...; F_M^{G^a}]$
- F^Gは最終的に全結合層へ $loss(y, \hat{y}) = \frac{1}{B} \sum_{i=1}^{B} y_i \log \hat{y}_i$
  - Lossはcross-entropy



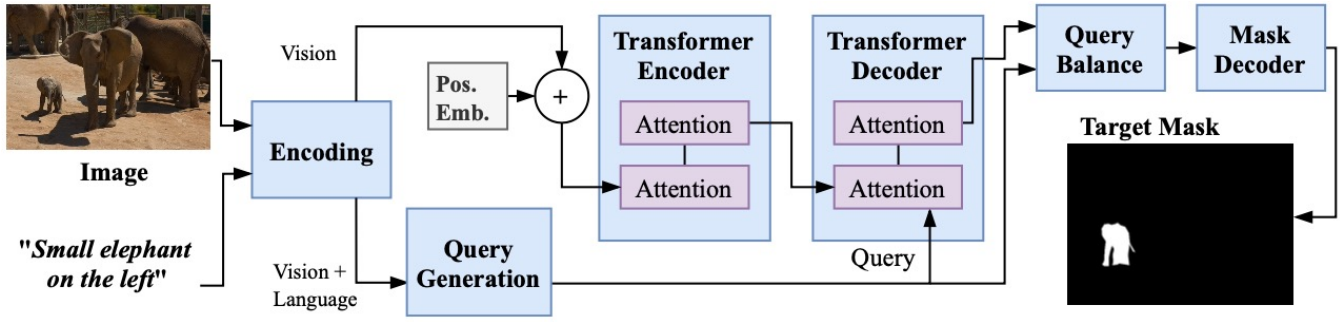


Islam, M. M., & Iqbal, T. (2020, October). Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 10285-10292). IEEE.

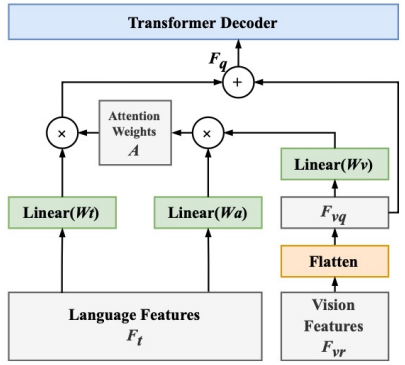
VLT [Ding+ (Nanyang Technological Univ), ICCV21]

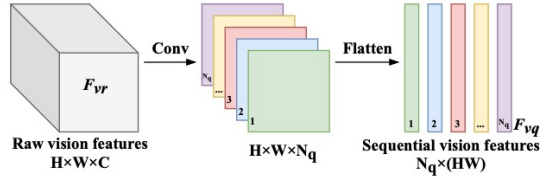
- Attention networkによるReferring Segmentation
  - 参照画像の助けを借りて参照文を理解するQuery Generation/Balance Moduleを提案
- Query Generation Module (QGM)
  - 画像特徴量F_{vq}、言語特徴量F_tを線形投射
    - $f_{vqn} \in R^{1\times(HW)}, n = 1, 2, \ldots, N_q$
    - $f_{ti} \in R^{1\times C}, i = 1, 2, \ldots, N_l$
  - $a_{ni} = \sigma(f_{vqn}W_v)\,\sigma(f_{ti}W_a)^T$

    $F_{qn} = A_n\sigma(F_tW_t)$
- Query Balance Module (QBM)
  - 各C_{qn}は、クエリF_{qn}が予測されたコンテキストにどれだけ適合するかを示す
- Mask Decoder
  - 3つの3×3 conv の後、1×1 conv でマスクを出力
  - 損失関数はマスクのBinary Cross Entropy



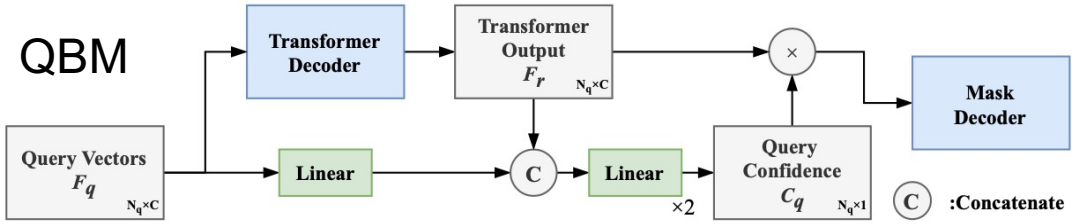
QGM





$$F_{vq} = \mathrm{Flatten}(\mathrm{Conv}(F_{vr}))^T$$

QBM



Ding, H., Liu, C., Wang, S., & Jiang, X. (2021). Vision-Language Transformer and Query Generation for Referring Segmentation. arXiv preprint arXiv:2108.05565.