Reference-based Image Super-Resolution with Deformable Attention Transformer

- https://arxiv.org/pdf/2207.11938

概要

- タスクはRefSR
- papers with codeでtop
- DATSR(Deformable Attention Transformer)を提案

related work

- RefSR
  - Robust reference-based super resolution via c2-matching(CVPR2021)
  - Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution(CVPR2021)
  - Towards content independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation(ECCV2020)
  - Feature representation matters: End-to-end learning for reference-based image super-resolution(ECCV2020)

RefSR

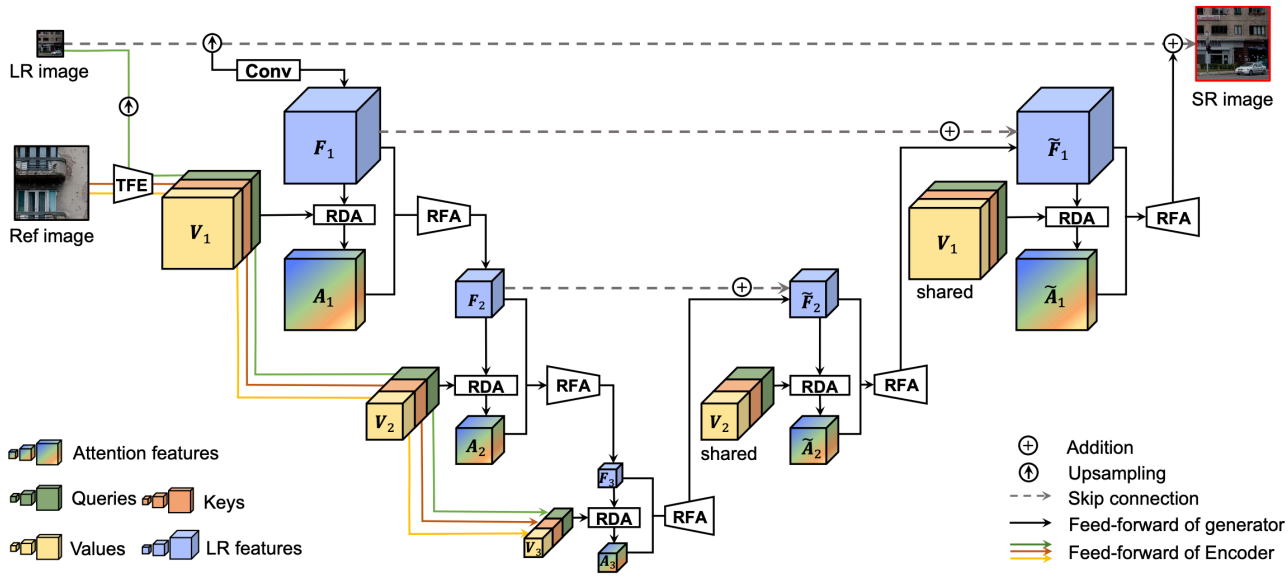
- Reference based Super Resolution
- LR1枚からHR1枚を生成するSISRとは異なりReference imageも入力にある
  - referenceの例とては別視点の高解像度画像など
- 以下の2つがchallenging
  - matching the correspondence between the LR and Ref images
    - 視点が異なるとき特に難しい
  - transferring textures of the high quality Ref images to restore the HR images
    - 関係ない情報も与えてしまうかもしれない

Multi-RefSR

- reference imageが複数あるRefSRをMulti-RefSRという
- model
  - CIMR-SR
  - AMRSR

DATSR

- 全体図は以下



- Texture Feature Encoders(TFE), Reference-based Deformable Attention(RDA), Residual Feature Aggregation(RFA)から構成される
  - TFE : extracting multi-scale texture features of Ref and LR images
  - RDA : matching the correspondences and transfer the textures from Ref images to LR images
  - RFA : aggregate features and generate SR images
- このTFE->RDA->RFAという処理の流れをU-Netの構造にして繰り返すようにしている

TFE

- I_{LR}をupscaleしてI_{Ref}と解像度をそろえたものを$I_{LR\uparrow}$とする
- このときencoderを通してq, k, vを得る

$$
\begin{aligned}
Q_l &= E_l^q\left(X_{LR\uparrow}\right) \\
K_l &= E_l^k\left(X_{Ref\uparrow}\right) \\
V_l &= E_l^v\left(X_{Ref\uparrow}\right)
\end{aligned} \tag{1}
$$

- 全体図のV_lとかかれている3つの立法体がQ, K, Vを表している

RDA

- 入力はTFEから得られたQ, K, VとI_{LR}のfeatureとして計算された前層の出力F_l
- 全体図は画像の通り

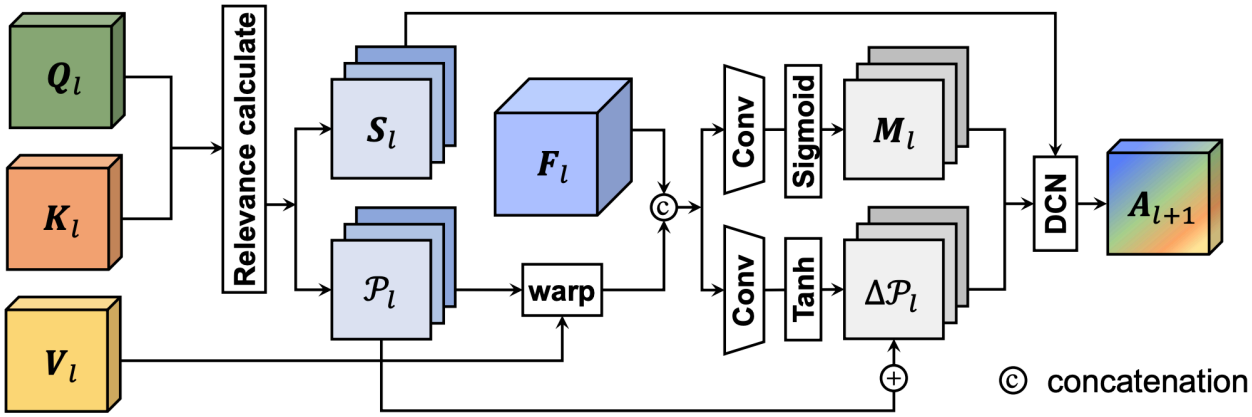


Fig. 3: The architecture of RDA.

- 数式で簡略的に記述すると以下

$$
\begin{aligned}
A_{l+1} &= \mathrm{RefAttention}\left(Q_l, K_l, V_l, F_l\right) \\
&= \mathcal{T}\left(\sigma\left(Q_l^{\top} K_l\right), V_l, F_l\right)
\end{aligned} \tag{2}
$$

- $\sigma\left(\cdot\right)$: correspondence matching function to calculate the relevance between the Ref and LR images
- $\mathcal{T}\left(\cdot\right)$: transferring the textures from the Ref to the LR image

RFA

- Swin Transformer Layerから構成される

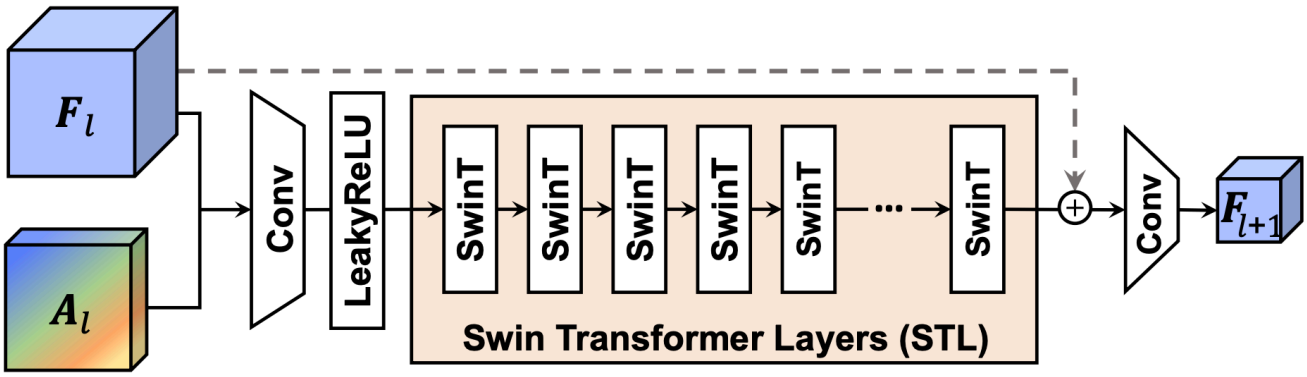


Fig. 4: The architecture of RFA.