# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | TransformerにおけるToken-Mixingの探索 |
| Title(English) | Exploring Token-Mixing Structure for Transformer |
| 著者(和文) | 浅倉 拓也, 宇都 有昭, 篠田 浩一 |
| Authors(English) | Takuya Asakura, Kuniaki Uto, Koichi Shinoda |
| 出典(和文) | 人工知能学会全国大会 (第36回)論文集, , , |
| Citation(English) | Proceedings of the Annual Conference of JSAI, , , |
| 発行日 / Pub. date | 2022, 6 |
| Note |  |

# Transformer における Token-Mixing の探索

Exploring Token-Mixing Structure for Transformer

浅倉 拓也　宇都 有昭　篠田 浩一
Takuya Asakura　Kuniaki Uto　Koichi Shinoda

東京工業大学 情報理工学院
School of Computing, Tokyo Institute of Technology

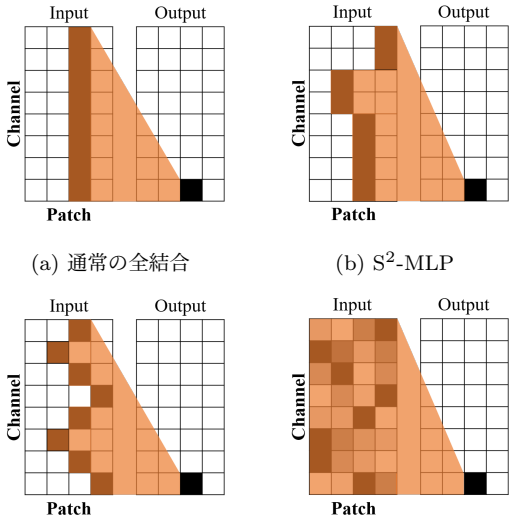
The Transformer model, which applies Channel-Mixing and Token-Mixing alternately to input data, has been developed for time-series data such as text and speech. Recent studies have shown that this model can also perform well image. Various improved models of transformers have been proposed for image processing, many of which have improved the structure of the fully connected layer, especially for Token-Mixing. However, these structures should be designed manually, which requires advanced knowledge about the characteristics of the target data. In this paper, we propose a method to automatically acquire Token-Mixing structures by learning the relationships between Tokens. In our experiments on the image classification tasks, the structure obtained by the proposed method achieves higher accuracy while having fewer parameters than the other Token-Mixing methods. We also visualized the Token-Mixing structures obtained by the proposed method, and observed that the proposed method tends to focus on spatially close Tokens.

## 1. はじめに

近年の深層学習による画像分類タスクにおいて，Transformer [Vaswani 17] をベースとした分類モデルが高い成果をあげており，従来の畳み込みニューラルネットワーク (CNN) の精度を超えつつある [Dosovitskiy 21]，[Yu 21]．これらは画像を複数のパッチに分割して扱い，異なるチャンネル同士，または異なるパッチ同士を学習可能な手法で繰り返し混合 (Mixing) することで特徴抽出が行われる．一般に，チャンネル方向の特徴抽出部は Channel-Mixing，空間方向の特徴抽出部は Token-Mixing と区別される．ViT[Dosovitskiy 21] や DeiT[Touvron 21] では，Transformer と同様に Channel-Mixing には全結合層を，Token-Mixing には Multi-head Attention(MHA) をそれぞれ用いている．しかしながら，MHA は自然言語処理において発展したものであり，画像処理における Token-Mixing として最適かどうかには疑問の余地があった．gMLP[Liu 21] では MHA をパッチ方向の全結合や要素積からなる Spatial Gating Unit(SGU) に置き換え，MHA を用いるモデルと同等以上の精度を達成した．$S^2$-MLP[Yu 22] では，MHA の代わりにパッチを空間方向に一つ分シフトしたのち全結合層を適用することで，空間的な特徴抽出を可能としている．これに類似した手法として，CycleMLP[Chen 22] では二つ以上離れたパッチ同士での全結合を行う CycleFC を提案している．また，[Yu 21] で提案された PoolFormer では Token-Mixing としてプーリングを用いており，単純な構造ながら他のモデルに比肩する精度を達成している．

画像処理モデルにおいて，Token-Mixing は空間的な特徴を抽出する目的で導入されているが，その空間的な特性は様々である．例として，MHA は全てのパッチ間，CycleFC は距離が離れた少数のパッチ同士，プーリングは近傍のパッチ同士で特徴抽出を行う．これらの応用を考えたとき，タスクに応じてどのような Token-Mixing が適しているのか判断するには深層学習と目的タスクに対する高度な知識が求められる一方，Token-Mixing の空間的特性と最終的なモデルの性能の関係は未だ不明瞭な部分が多い．さらに，これらのモデルは基本的に 1 種類の Token-Mixing を導入しているため，そのモデルの層や構造ごとに最適な空間的特性を持つとは限らず，それらを網羅的に検証することは計算コストの面で現実的ではない．

そこで，本研究では Token-Mixing に対して多様な空間的特性を表現可能な探索空間を設定し，最適な条件を探索する手法，Adaptive Fully Connected Layer(AdaFC) を提案する．Token-Mixing に AdaFC を適用し，その他の部分を従来手法と同様にしたモデルを用いて ImageNet[Russakovsky 15] の分類タスクを学習し，正答率を比較することで評価を行った．また，得られた Token-Mixing を可視化することで，層の深さによって異なる空間的特性を持ち，浅い層では近傍のパッチ同士で，深い層では離れたパッチ同士でそれぞれ特徴抽出を行うことが確認された．




(a) 通常の全結合　(b) $S^2$-MLP

(c) CycleMLP　(d) AdaFC で得られる一例

図 1: 全結合層を利用した Token-Mixing の空間的特性

連絡先: 浅倉拓也，東京工業大学 情報理工学院，
asakura@ks.c.titech.ac.jp

## 2. 提案手法

### 2.1 Adaptive Fully Connected Layer

パッチ間の特徴抽出において，CycleMLP や $S^2$-MLP では全結合層を応用した手法を提案している．通常の全結合の特徴抽出は図 1 (a) のように，入力側の全てのチャンネルにおいて同じパッチの値を抜き出して重み付けする．ここで，左右のグリッドはそれぞれ入力および出力特徴量であり，図中のオレンジ色の領域が全結合による重み付けと加算を表している．$S^2$-MLP では，入力特徴量をパッチ方向にシフトして全結合を行うため，図 1 (b) に示すように異なるパッチ同士を抜き出して重み付けすることができる．CycleMLP では重みづけするパッチ位置について図 1 (c) のような一定のパターンを用意することで，より複雑な位置同士の特徴抽出を実現している．このようにチャンネル方向の全結合を応用することで，入力画像の大きさに制限がなくなり，パッチサイズが増加した場合でもパラメータ数を大きく増やすことなく Token-Mixing を実装できるという利点がある．一方で，これらの手法で特徴抽出できる距離は層によらず一定であることや，入力側の各チャンネルと抜き出すパッチが常に一対一の関係でなければならないといった問題もある．

本研究では，図 1 (d) に示すような，複数のパッチを学習可能な比重で抜き出して各チャンネルと対応付けたのち重み付けする AdaFC を実装し，この比重のパターンを最適化することを目的とする．AdaFC は図 2 のように表される．また，仮想コードをアルゴリズム 1 に示す．$P$ パッチからなる $C$ チャンネルの入力特徴量 $\boldsymbol{X} \in \mathbb{R}^{C\times P}$ が与えられるとき，学習可能なパラメータ $\boldsymbol{W} \in \mathbb{R}^{C\times P}$ を用意する．本提案手法ではパラメータ $\boldsymbol{W}$ はあるパッチ $p$ における全パッチとの重要度をチャンネルごとに表現した重み行列と捉え，$\boldsymbol{W}$ に対してソフトマックス関数をかけ，これを適用するために $\boldsymbol{X}$ と要素積を計算したのち，和を求めて長さ $C$ のベクトル $\boldsymbol{y_p}$ を得る．このとき，$\boldsymbol{W}$ が最適化されるに従い任意のパッチ $p$ に対して関連度が高いパッチを抽出することが期待される．$\boldsymbol{W}$ をパッチ方向にシフトしながら全てのパッチに対してこの操作を繰り返し，全ての $\boldsymbol{y_p}$ を結合することで全てのパッチに対して $\boldsymbol{W}$ を適用した出力値 $\boldsymbol{Y} \in \mathbb{R}^{C\times P}$ が得られる．最後に得られた $\boldsymbol{Y} \in \mathbb{R}^{C\times P}$ に対して一般的な全結合層での線形変換を行うことで，異なるパッチ間での特徴抽出が可能となる．

### 2.2 モデル構成

本研究では従来研究と精度を比較することで提案手法の評価を行うため，Token-Mixing の部分を除いて [Yu 21] と同様の構造を持つモデルを用いる．このモデルは四つのステージからなり，各ステージでは最初に画像または入力特徴量をカーネルサイズ $3\times 3$，ストライド 2 の畳み込み層に通す．これは 1 番目のステージにおいて画像をパッチ化し，それ以降では特徴量を圧縮し学習を効率化する効果がある．各ステージでは図 3 に示すブロックを繰り返し適用して特徴抽出を行う．また，正規化層には Group Normalization を用いる．

パッチ化された画像には高さ及び幅が存在するため，縦方向と横方向それぞれについて重みパラメータ $\boldsymbol{W}$ を用意し，それらを適用したのち全結合を行う．厳密には，縦方向と横方向の適用順によって得られる特徴に違いが生じるものの，それらは学習によって吸収できるものと判断した．$\boldsymbol{W}$ は各 Token-Mixing に対して用意することも可能ではあるが，パラメータ数が増大し学習が長期化するため，ステージごとに最適化し，それぞれのステージ内では同じものを使用する．よって，学習モデルは縦方向および横方向にそれぞれ四つの $\boldsymbol{W}$ を持つ．

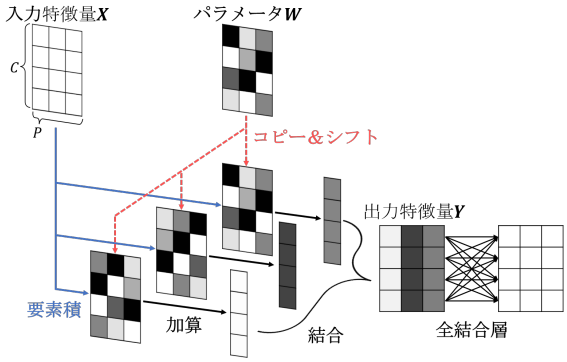



図 2: AdaFC による Token-Mixing

**アルゴリズム 1** AdaFC における重み適用の仮想コード

**Input:** $\boldsymbol{X} \in \mathbb{R}^{C\times P}$, $\boldsymbol{W} \in \mathbb{R}^{C\times P}$, $C > 0$, $P > 0$
**Output:** $\boldsymbol{Y} \in \mathbb{R}^{C\times P}$

```
for (c = 0; c < C; c++) do
    W[c, :] ← Softmax(W[c, :])
end for
for (p = 0; p < P; p++) do
    Y[:, p] ← Sum(X ⊙ W, dim = 1)
    W[:, 1 :] ← W[:, : P − 1]
    W[:, 0] ← W[:, P − 1]
end for
```

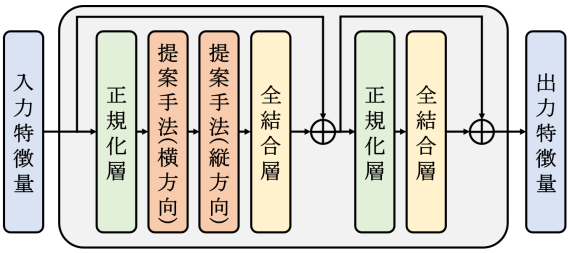



図 3: Token-Mixing に提案手法を用いたブロック

## 3. 実験内容

Token-Mixing として AdaFC を採用したモデルを用いて，ImageNet-1K[Russakovsky 15] の 1000 クラス分類タスクを学習させた．ImageNet の画像サイズは 224pixel であるため，ステージ 1 から 4 のパッチサイズはそれぞれ 112，56，28，14 となる．実験で用いるモデルの $s$ 番目のステージにおけるチャンネル数 $C_s$ は $[C_1, C_2, C_3, C_4] = [64, 128, 320, 512]$ とした．本研究では，$s$ 番目のステージにおけるブロックの繰り返し数 $N_s$ を $[N_1, N_2, N_3, N_4] = [2, 2, 2, 2]$ としたモデル-S，$[N_1, N_2, N_3, N_4] = [2, 2, 6, 2]$ としたモデル-M，$[N_1, N_2, N_3, N_4] = [4, 4, 12, 4]$ としたモデル-L の三つを作成して学習させた．比較を容易にするため，ハイパーパラメータ及び正則化の大部分について [Yu 21] と同じものを設定した．最適化アルゴリズムは AdamW とし，バッチサイズは 2048 として 300Epoch として学習させた．学習率は Warmup 時に $10^{-6}$ から $10^{-3}$ まで 5Epoch かけて増加させ，それ以降

表 1: 他のモデルとの ImageNet における分類精度の比較

| モデル | パラメータ数 | 正答率 (%) |
|---|---|---|
| ViT-B/16 [Dosovitskiy 21] | 86M | 79.7 |
| DeiT-S [Touvron 21] | 22M | 79.8 |
| gMLP-Ti [Liu 21] | 6M | 72.3 |
| gMLP-S [Liu 21] | 20M | 79.6 |
| $S^2$-MLP-wide [Yu 22] | 51M | 80.7 |
| CycleMLP-B1 [Chen 22] | 15M | 79.1 |
| CycleMLP-B2 [Chen 22] | 27M | 81.6 |
| PoolFormer-S12[Yu 21] | 12M | 77.2 |
| PoolFormer-S24[Yu 21] | 21M | 80.3 |
| モデル-Init | 13M | 76.5 |
| モデル-S | 9M | 76.4 |
| モデル-M | 13M | 79.4 |
| モデル-L | 24M | 81.1 |

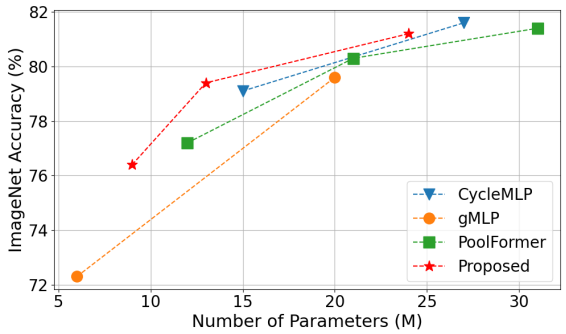
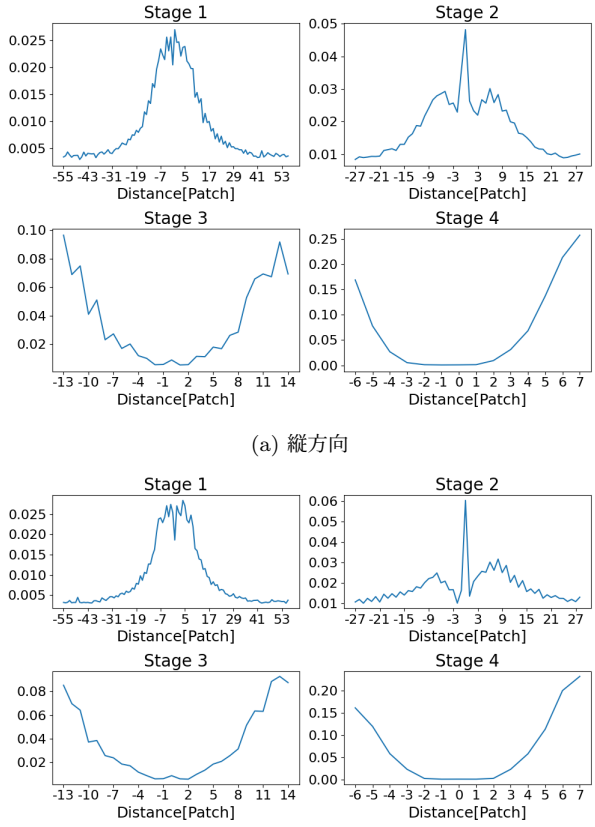



図 4: パラメータ数と正答率の比較




(a) 縦方向

(b) 横方向

図 5: AdaFC によるパッチ間の関連度合いの可視化

は最大値 $10^{-3}$，最低値 $10^{-6}$ の半周期の Cosine 関数に従って減少させた．また，正則化のため Data Augmentation に AutoAugment を用い，0.05 の Weight Decay，0.01 の Label Smoothing，Mixup または CutMix，確率的に層の数を減らす Stochastic Depth を適用した．

AdaFC の重みパラメータ $\boldsymbol{W}$ は平均 0，分散 $10^{-2}$ の正規分布で初期化した．実験では $\boldsymbol{W}$ を最適化することの有効性も同時に確かめるため，モデル-M の $\boldsymbol{W}$ を初期値で固定したまま他のパラメータのみ最適化を行うモデル-Init を作成し，精度の比較対象に加えた．また，$\boldsymbol{W}$ を導入し学習することで他の Token-Mixing よりも必要となる計算時間が増加することが予想できる．本研究では Token-Mixing にプーリング層または通常の全結合層を適用したモデルを作成し，AdaFC と計算時間の比較を行った．

## 4. 結果と考察

### 4.1 精度および計算時間の比較

モデル-S，モデル-M，モデル-L およびモデル-Init のパラメータ数と学習後の正答率を表 1 に示す．また，パラメータ数を横軸に，正答率を縦軸にとったグラフを図 4 に示す．モデル-M は PoolFormer-S12 と，モデル-L は PoolFormer-S24 と Token-Mixing の部分を除いて全く同じ構造であり，またパラメータ数もほぼ同じでありながら，どちらについても AdaFC を用いたモデルが高い正答率を達成している．さらに，モデル-L のパラメータ数は $S^2$-MLP の半分以下であるにも関わらず，精度に関しては上回っている．また，モデル-Init の正答率はモデル-M と比べて大きな差があり，PoolFormer-S12 よりも低くなってしまっている．これらのことから，$\boldsymbol{W}$ の最適化の有無によって正答率が大きく変わり，提案手法によって複数のパッチ同士での特徴抽出を行いながら，そのパターンを学習することの有効性が示されている．

CycleMLP と比較した場合，モデル-M については CycleMLP-B1 の精度を上回っている一方，モデル-L の精度は CycleMLP-B2 よりも低くなっている．これには CycleMLP-B2 と比べてパラメータ数が少ないことや，Token-Mixing 以外の構造が完全には一致していないことも影響しているものの，最大の要因は AdaFC が過学習しやすいことにあると考えられる．そのため，今後の精度向上のためには正則化の手法について調整が必要である．

1Epoch あたりの学習にかかる時間について比較すると，Token-Mixing を AdaFC を使用したモデルは全結合層とした場合より 11.1%，プーリング層とした場合より 13.5%増加する結果となった．

### 4.2 可視化

AdaFC の重みパラメータ $\boldsymbol{W}$ を確認することで，特徴抽出に利用したパッチの距離や比重を知ることができる．学習後のモデル-L について，各ステージで使用した $\boldsymbol{W}$ を取り出し，チャンネル方向に平均をとることでステージごとの Token-Mixing の空間的特性を可視化する．縦方向について図 5 (a) に，横方向について図 5 (b) にそれぞれ示す．中央付近の値が高いほど近傍のパッチ同士で，左右の値が高いほど離れたパッチ同士で特徴抽出していることを表している．すなわち，通常の全結合

層であれば中央にのみピークが立つ．

結果から，全体的な傾向として層の浅い部分では近傍同士の特徴抽出を行い，層が深くなるに従って離れた位置同士を関連付けていることが分かる．ステージ 1 でのパラメータは正規分布に従うように近傍同士を関連付けている一方，ステージ 2 では近傍から少し離れた距離に着目している．また，ステージ 1 では距離が 0 の時の値，すなわち同じパッチ同士での関連度が少し低くなっているのに対し，それを補うようにステージ 2 では最も値が高くなっている．学習された空間的特性はステージからステージ 2，ステージ 3 からステージ 4 にかけてそれぞれ連続的に変化している一方，ステージ 2 からステージ 3 にかけては急激に変化している．モデル-L のステージ $s$ で繰り返されるブロック数 $N_s$ は $[N_1, N_2, N_3, N_4] = [4, 4, 12, 4]$ であり，ブロック数が急激に増加するステージ 2 からステージ 3 において，不連続な変化が生じたと考えられる．

CNN では畳み込みを用いた特徴抽出と圧縮を重ねることで，小さい領域から広範囲へと次第に受容野を広げており，AdaFC の学習も同様の傾向に収束したものと考えられる．一方で，近傍同士の特徴抽出のみを重ねた場合，最終的な出力に対する離れた位置同士の重要性は薄れてしまうことが知られている [Luo 16]．対して，AdaFC では層の深い領域で離れた位置同士を関連付けるように学習できているため，入力画像全体から満遍なく特徴抽出が可能であると考えられる．このように，AdaFC では効果的な Token-Mixing を学習できるだけでなく，層の深さに応じて適した空間的特性を獲得することができた．

## 5. まとめと今後の課題

本研究では，Transformer における最適な空間的特性を持つ Token-Mixing を獲得可能な AdaFC を提案した．従来研究で提案されたモデルに AdaFC を取り入れて学習させることで，他の Token-Mixing よりも高い精度を達成することができた．また，学習後の重みを可視化することで，AdaFC が層ごとに特性のことなる Token-Mixing を学習可能であり，低い層では画像の近傍間，深い層では遠い位置同士の特徴抽出が有効であるという知見が得られた．一方，AdaFC は過学習を起こしやすく，特にパラメータ数が大きいモデルについては正則化の調整が必要である．

今回は対象とするタスクを画像分類に絞って提案手法を評価したが，今後は物体認識やセグメンテーションなど別のタスクを目的とした場合や，扱うデータを画像から音声やテキストに変えた場合にどのような空間的特性に収束するのか検証していきたい．加えて，得られた重みの定量的な評価方法についても検討していきたい．

## 謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP18002）の結果得られたものです．

## 参考文献

[Chen 22] Chen, S., Xie, E., GE, C., Chen, R., Liang, D., and Luo, P.: CycleMLP: A MLP-like Architecture for Dense Prediction, in *International Conference on Learning Representations* (2022)

[Dosovitskiy 21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *ICLR* (2021)

[Liu 21] Liu, H., Dai, Z., So, D., and Le, Q. V.: Pay Attention to MLPs, in Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. eds., *Advances in Neural Information Processing Systems* (2021)

[Luo 16] Luo, W., Li, Y., Urtasun, R., and Zemel, R.: Understanding the Effective Receptive Field in Deep Convolutional Neural Networks, in Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. eds., *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc. (2016)

[Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)*, Vol. 115, No. 3, pp. 211–252 (2015)

[Touvron 21] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H.: Training data-efficient image transformers & distillation through attention, in *International Conference on Machine Learning*, Vol. 139, pp. 10347–10357 (2021)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I.: Attention is All you Need, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. eds., *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. (2017)

[Yu 21] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S.: MetaFormer is Actually What You Need for Vision, *arXiv preprint arXiv:2111.11418* (2021)

[Yu 22] Yu, T., Li, X., Cai, Y., Sun, M., and Li, P.: S2-MLP: Spatial-Shift MLP Architecture for Vision, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 297–306 (2022)