画像変換を用いたゼロショットドメイン適応

中西　優司†　　白井　匡人††

† 島根大学 自然科学研究科　〒 690–0823 島根県松江市西川津町 1060
†† 島根大学 学術研究院理工学系　〒 690–0823 島根県松江市西川津町 1060
E-mail: †n21m108@matsu.shimane-u.ac.jp, ††shirai@cis.shimane-u.ac.jp

あらまし　本論文では，画像分類タスクにおいて学習データの無いクラスを分類するために Image-to-Image Translation を用いて学習データを生成する手法を提案する．具体的には画像分類を行いたいデータセットと同じクラスを持つ別のデータセットを基に画像変換モデルを構築することで，対象のデータセットの一部のクラスの学習データが存在しない場合に学習データの無いクラスの疑似的な学習データを生成する．実験では一部のクラスの学習データを Image-to-Image Translation により生成した画像としたときの分類性能を検証する．

キーワード　画像分類，Zero-shot Domain Adaptation，Image-to-Image Translation

704704

1 前　書　き

近年，機械学習の発展により，画像分類やオブジェクト検出など様々な分野で多くの研究が行われいる．これらは非常に精度が高くなってきており，その根底にあるのは膨大な学習データの存在である．学習データが少ない，または質が悪いと，モデルの性能に大きく影響してしまう．

学習データが少ない場合は，元の学習データに回転や移動などの変換を加えてデータ数を水増しする手法や，敵対的手法によってドメイン共通の特徴を捉えて本物の学習データに近い擬似学習データを生成する手法など，Data Augmentation と呼ばれる研究も盛んに行われている．しかし，実際に適用する際に，特定のクラスの学習データが全くない場合が存在する．

このような場合には，ゼロショット学習という手法も適用することができる．ゼロショット学習は，画像のほかに色や羽の有無などの属性やテキスト情報を補助情報として利用することで補助情報のみしか利用できない未知のクラスの画像を分類することを目的とした手法である．しかし，このような手法は，補助情報に多くのノイズが含まれるために大規模なデータセットを使用しない限り汎用性の低いものとなる．分類したいデータセットに学習データが存在しない場合，画像変換の学習データとして対応する別のデータセットが存在すると仮定してゼロショットドメイン適応手法が適用できる．ゼロショットドメイン適応では，分類したいデータセットと同じクラスを持つ別のデータセット，分類したいデータセットと色や背景などの表現の対応を持つ 2 つの別のデータセットをそれぞれ利用し，画像分類のための学習データを生成する．3 つのデータセットそれぞれに十分な学習データがあることが必要であるため，分類したいデータセットの学習データが部分的に存在し，学習データが存在しないクラスが増減する本研究に適用することは困難である．

本稿では，画像分類の学習データとして利用するために 1 つのノイズや画像から多様なデータを生成可能な Image-to-Image Translation に着目する．Image-to-Image Translation は複数の種類のデータセットを用いて学習し，一方のデータセットの画像の位置や向きなどの特徴を維持したままもう一方のデータセットの表現を合成することができる．しかし，本研究では 2 つのデータセットから複数のクラスをラベルなしで学習データとして利用するため，表現の学習が困難になる場合がある．そのため本稿では，位置や向きなどの特徴が近い画像ペアを学習する際に重みが大きくなるようにすることで，複数のクラスを用いた Image-to-Image Translation においても正しく表現が合成される手法を提案する．

第 2 章では関連研究について述べ，第 3 章では，提案手法について述べる．第 4 章では実験により，提案手法が欠落したクラスの多様なデータを生成することができるか検証を行い，学習データがないクラスの画像として生成したデータを利用した際の比較を行う．第 5 章で結論とする．

2 関 連 研 究

2.1 ゼロショット学習

ゼロショット学習では，学習データに含まれない未知のクラスのみ，または未知のクラスを含むテストデータを分類可能にすることを目的としており，特徴空間のみを用いる通常の画像分類とは異なり，未知のクラスを識別するための情報を埋め込むための意味空間が存在する．具体的には，画像と同時に，補助的な情報として色などの色覚的な特徴やくちばしや手足の有無といった形状的な特徴などの様々な属性情報 [1] を数値化して学習したり，画像やクラスを表すテキスト情報 [2] を単語埋め込みモデルを利用して数値化し，学習する．しかし，これらの情報には多くのノイズが存在する問題や，ノイズ低減のために膨大な学習データが必要であるという問題が存在する．

2.2 ゼロショットドメイン適応

ゼロショットドメイン適応は Peng [3] らによって提案された

図 1　ゼロショットドメイン適応において数字データセットと衣類データセットを用いるときの定義

ドメイン適応タスクであり，分類したいデータセットを全く用いることなく，ドメイン適応によって生成した画像のみを学習データとして用いて画像分類を行う．分類したいデータセットと同じクラスを持つ別のデータセット，分類したいデータセットと色や背景などの表現の対応を持つが，先に挙げた 2 つとは分類タスクの異なる別のデータセットを学習データとして利用する．例としてカラーの数字データセットである MNIST-M [12] のクラス分類を行う場合に，グレースケールの数字データセットである MNIST [11] と，タスクの異なる 2 つのデータセットとして衣類のグレースケールデータセットである FashionMNIST [13] と FashionMNIST のカラーデータセットを用いるとき，図 1 のように表すことができる．Wang らによって提案された Conditional Coupled GAN [4] では，Coupled GAN [7] のモデルを用いてゼロショットドメイン適応に拡張している．Coupled GAN は中間層を共有する 2 つの生成器と出力層を共有する 2 つの識別器を用いてそれぞれの領域の画像を学習させ，1 つのランダムノイズを，中間層の共有された生成器に入力することで，一対一の対応関係を持つ両領域の画像ペアを生成することができる手法である．図 1 のデータセットを用いる場合，グレースケールの 2 つのデータセットを情報源領域，カラーのデータセットを対象領域として 2 組の Coupled GAN モデルに学習させ，衣類と数字のどちらを生成するか決定するバイナリ変数を用いることによってグレースケールとカラー，衣類と数字を組み合わせた 4 つのデータセットの疑似的な画像を生成することが可能になる．生成した数字のカラー画像を画像分類に利用することで MNIST-M の画像を用いることなくドメイン適応を行う．

2.3　Image-to-Image Translation

Image-to-Image Translation では，2 つの領域の画像において，一方の画像またはノイズを入力としてオブジェクトの位置や向きなどを維持してもう一方の領域の画像を生成することが可能である．Image-to-Image Translation は GAN [5] に基づいて生成器と識別器を利用している．両領域に対する 2 つの生成器を用いて変換された入力画像を再度変換した差を損失として与えることで一対一の変換を行う CycleGAN [6]，2 つの生成器のパラメータを共有することで同じノイズから両領域で一対一の関係

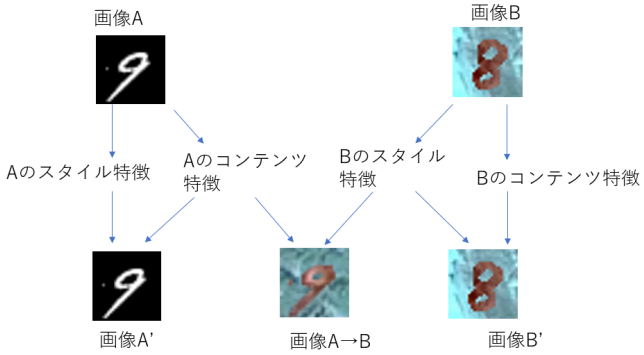

図 2　コンテンツ特徴とスタイル特徴
領域 X の画像 A と領域 Y の画像 B から抽出したコンテンツ特徴とスタイル特徴を利用して A から領域 Y の画像を生成可能．

を持つ画像を生成可能な CoupledGAN [7]，いくつかのエンコーダを用いて，学習データから位置や向きなどの構図を表すコンテンツ特徴と見た目や背景などを表すスタイル特徴に分離する手法 [8] [9] などが提案されている．分離した特徴は 2 のように組み合わせることで領域間の変換が可能になる．GP-UNIT を含む Image-to-Image Translation 手法では，本研究のように複数のクラスがテストデータとして与えられ，テスト時に入力される画像のクラスが学習データに含まれていない場合，正しくスタイル特徴を抽出することが困難な場合があると考えられる．本研究では，この問題を改善するため，GP-UNIT の損失関数を調整するパラメータを追加することを提案する．

3　提 案 手 法

3.1　設　　定

画像分類を行いたい対象領域のデータセットを Y，もう一方の情報源領域のデータセットを X とし，Y にてデータが存在するクラスを $(Y_{s1}, Y_{s2}... \in Y_s,\ X_{s1}, X_{s2}... \in X_s)$，データが存在しないクラスを $(Y_{u1}, Y_{u2}... \in Y_u,\ X_{u1}, X_{u2}... \in X_u)$ とする．ここで，X_u には学習データが存在するものとする．学習時には Y_s，X_s をクラスラベルなしで利用し，学習後の入力として X_u のクラスの画像を利用する．

3.2　コンテンツエンコーダの学習

本手法は GP-UNIT [9] に従って二段階で構成し，第一段階の学習を次のように行う．GP-UNIT では BigGAN [10] などの条件付き GAN が同じ入力ノイズから複数の領域の画像を生成できることに着目し，同じ入力ノイズ生成した生成した画像をペアとしてコンテンツ特徴と呼ばれる位置や向きなどから抽出される特徴が一致するようエンコーダを学習させる．具体的には，BigGAN にてランダムノイズを z_1，z_2，z_3... としたとき，入力ノイズ z_1 と，領域 X のラベル l_x，領域 Y のラベル l_y から生成された画像を同じコンテンツ特徴を持つ画像ペア $(x_1,\ y_1)$ として，コンテンツ特徴を抽出するコンテンツエンコーダ E_{con}，スタイル特徴を抽出するスタイルエンコーダ E_{sty}，デコーダ F を

以下の損失で学習させる．

コンテンツエンコーダ E_{con}，スタイルエンコーダ E_{sty} を用いて抽出した特徴を領域 X のドメインラベル l_x と共にデコーダ F に入力し，再構成された画像と元の画像との L2 距離によって再構成損失を定義する．

$$L_{rec} = \mathbb{E}_x[\|x_1 - F(E_{con}(x_1), E_{sty}(x_1), l_x)\|_2] \tag{1}$$

コンテンツエンコーダにより抽出された x_1 のコンテンツ特徴を入力したときのデコーダ F の中間層 F_{sty} の出力と x_1 のスタイル特徴の L1 距離によってスタイル再構成損失を決定する．

$$L_{rsty} = \lambda_{sty}\mathbb{E}_x[\|F_{sty}(E_{con}((x_1), l_x) - x_{sty})\|_1] \tag{2}$$

ここで，λ_{sty} はハイパーパラメータとする．次に，同じ入力ノイズから BigGAN によって生成された x_1 と y_1 は同じコンテンツ特徴を持つことを前提としているため，x_1 と y_1 をコンテンツエンコーダ E_{con} に入力したときの出力を比較する．また，y_1 のコンテンツ特徴と領域 X のドメインラベル l_x を入力したときのデコーダ F の中間層 F_{sty} の出力から得たスタイル特徴と，x_1 のスタイル特徴 x_{sty} が一致するよう損失を与える．

$$L_{rcon} = \mathbb{E}_{x,y}[\|(E_{con}(x_1) - E_{con}(y_1)\|_1 + \lambda_{sty}\|F_{sty}(E_{con}(y_1), l_x) - x_{sty})\|_1] \tag{3}$$

コンテンツ特徴は領域に依存しないため，領域分類器 C を用いて次のような損失を定義する．

$$L_{dg} = \mathbb{E}_x[-logC(E_{con}(x_1))] + \lambda_r\mathbb{E}_x[E_{con}(x_1)] \tag{4}$$

ここで，λ_r はハイパーパラメータとする．したがって，第一段階の損失は次のようになる．

$$\min_{E_{con},E_{sty},F,C} = L_{rec} + L_{rsty} + L_{rcon} + L_{dg} \tag{5}$$

第二段階では，第一段階で学習させたコンテンツエンコーダ E_{con} を固定してスタイルエンコーダ E_{sty} を含む生成器 G，識別器 D の学習を行う．

最初に，領域 X の画像 I_x と領域 Y の画像 I_y からコンテンツエンコーダ E_{con} とスタイルエンコーダ E_{sty} を用いて特徴 x_{con}，y_{con}，y_{sty} を抽出し，生成器 G の入力とする．x_{con}，y_{sty} を入力として生成された I_{xy} は識別器 D によって領域 Y の画像か，そうでないか識別される．

$$L_{ad} = \mathbb{E}_y[logD(I_y)] + \mathbb{E}_{x,y}[log(1 - D(I_{xy}))] \tag{6}$$

D の中間層 D_{sty} を用いて I_{xy} と I_y のスタイル特徴を抽出する．

$$L_{sty} = \mathbb{E}_{x,y}[\|(D_{sty}(I_{xy}) - D_{sty}(I_y))\|_1] \tag{7}$$

また，コンテンツエンコーダ E_{con} によって I_{xy} と I_x のコンテンツ特徴が抽出され，2 つの画像から同様のコンテンツ特徴が得られるように損失が与えられる．ここで，コンテンツエンコーダは固定されているため，学習の中で一貫したコンテンツ特徴の抽出が可能である．

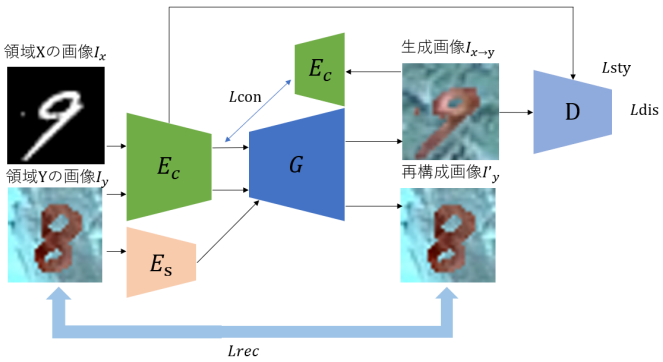

図 3　提 案 手 法

$$L_{con} = \mathbb{E}_{x,y}[\|E_{con}(I_{xy}) - E_{con}(I_x)\|_1] \tag{8}$$

y_{con}，y_{sty} を生成器 G の入力として生成した画像 I'_y と元の領域 Y の画像 I_y を比較し，I_y に近い画像が再構成されるよう損失を与える．

$$L_{rec} = \mathbb{E}_y[\|I'_y - I_y\|_1] \tag{9}$$

3.3　学習データが存在しないクラスの Image-to-Image Translation

本手法では，Image-to-Image Translation において学習データにテスト時の入力となる画像のクラスが含まれていない場合に対処する．学習済みコンテンツエンコーダは BigGAN によって生成された異なる領域とクラスの画像から学習しており，クラスに囚われないコンテンツ特徴，スタイル特徴の抽出が可能であると考える．生成器，スタイルエンコーダについては学習データに複数のクラスが含まれることによる影響を受けると考え，スタイル損失がコンテンツ特徴の近い画像に対して重点的に学習されるよう式 (7) を変更する．図 3 に提案手法のモデル図を示す．一組のバッチを学習する度に，コンテンツエンコーダ E_{con} を用いて I_x と I_y のコンテンツ特徴 x_{con}，y_{con} を抽出し，保持する．D の中間層 D_{sty} を用いて I_{xy} と I_y のスタイル特徴を抽出し，同様の出力が得られるよう損失を与える．この際，バッチ内の画像のコンテンツ特徴 x_{con}，y_{con} の L1 距離を計算し，それによって損失の重みを決定する．x_{con}，y_{con} の L1 距離は 0 から 1 に正規化されている．

$$L_{csty} = \mathbb{E}_{x,y}[\|(D_{sty}(I_{xy}) - D_{sty}(I_y))\|_1 \times (1 - \alpha)] \tag{10}$$

$$\alpha = \|E_{con}(I_y) - E_{con}(I_x)\|_1 \tag{11}$$

最終的な損失は以下のようになる．

$$\min_{G,E_{sty}} \max_{D} = L_{ad} + \lambda_{csty}L_{csty} + \lambda_{con}L_{con} + \lambda_{rec}L_{rec} \tag{12}$$

ここで，λ_{csty}，λ_{con}，λ_{rec} はそれぞれの損失のバランスをとるためのハイパーパラメータである．

4 実　　験

4.1 データセット

本研究では，MNIST [11] と MNIST-M [12] の，2 つのデータセットを用いて実験を行う．MNIST は，クラス 0 から 9 で，それぞれ 5923 枚，6742 枚，5928 枚，6131 枚，5842 枚，5421 枚，5918 枚，6265 枚，5851 枚，5949 枚の計 60000 枚の学習データと 10000 枚のテストデータを含む手書き数字データセットである．これらのデータは完全にラベル付けされている．MNIST-M は MNIST の数字に加えて，背景として BSDS500 [14] からランダムに抽出されたパッチを貼り付けたデータセットであり，すべてラベル付けされている．

4.2 提案手法を用いた画像生成

MNIST-M の Y_u の数と内容クラスを変化させて，提案手法と MUNIT，GP-UNIT により学習データにないクラスの画像を Image-to-Image Translation により生成する．MNIST と MNIST-M の両データセットにおいて，学習時には X_s，Y_s を利用し，学習後のモデルに X_u のクラスの画像を入力して Y_u の画像を生成する．MNIST と MNIST-M の 2 つのデータセットには一対一の対応関係が存在するが，学習時は対応関係とクラスラベルを利用せず，学習後に X_u の画像を一クラスずつランダムに入力して生成画像のクラスラベルを得る．MUNIT のバッチ数は 1，epoch 数は 10 とし，GP-UNIT，提案手法のバッチ数は 2，iteration 数は 75000 とする．学習後に，X_u からランダムに選んだ画像で構成したテストデータをコンテンツ特徴を抽出する画像として入力し，スタイル特徴を抽出する画像として Image-to-Image Translation の学習に用いた画像をランダムに入力する．Y_u を 9 と 7，8，9 としたときの生成結果をそれぞれ図 4，5 に示す．

図 4　9 を未知クラスとしたときの生成結果

4.3 生成データを用いた分類結果

4．2 と同様の方法で画像分類タスクの学習データを生成し，MNIST-M の Y_u のクラスの画像と置き換える．Y_u のクラスを 9，8 と 9，7 と 8 と 9 としたときの 3 つの設定において，提案手法と MUNIT，GP-UNIT により学習データを生成して利用した場合の分類精度を比較する．テストデータは各クラス 1000 枚ずつの計 10000 枚，モデルは InceptionV3 [15], ResNet50 [16] を使用する．これらの分類モデルは共に ImageNet [17] で事前学習が行われているものを利用し，バッチサイズは 32，epoch 数は 20 とする．以下に X_s のクラス数や種類を変更して提案手法と MUNIT，GP-UNIT 生成した画像で学習データを代替した場合の InceptionV3 と ResNet50 での F1-score を示す．

図 5　7，8，9 を未知クラスとしたときの生成結果

4.4 考　　察

図 4 より，未知クラスを 9 のみとしたときはすべての手法が識別可能な画像を生成しているが，MUNIT では文字の色がやや単調になっているのに対し，GP-UNIT，提案手法は多様な文字の色と背景が合成できている．また，図 5 より，Y_u のクラスの数を増加させたとき，MUNIT では文字の周りが白っぽくなっていて，文字の色が背景と同化している部分も確認できる．対して，GP-UNIT と提案手法では Y_u を 9 のみとした場合と変わらず質の高い画像が生成できている．表 1，2 より，Y_u を 1 クラス，2 クラスしたときや既存手法の方が優れている結果であるが，7，8，9 と 3 クラスにした場合は既存手法は大きく精度が低下する．InceptionV3 では 2 番目に高い精度を示した GP-UNIT を 0.055 改善し，ResNet50 では 0.012 改善することができている．また，Y_u を 8，9 としたとき，InceptionV3 では MUNIT，ResNet50 では GP-UNIT を利用して学習データを生成した場合が最も精度が高くなっている．しかし，既存手法は 2 つの分類モデルによって精度に大きな差があるが，提案手法では安定した精度を実現し，InceptionV3 と ResNet50 での結果を平均すると MUNIT を利用して学習データを生成した場合が 0.907，GP-UNIT を利用して学習データを生成した場合が 0.910，提案手法を利用して学習データを生成した場合が 0.920 であり，提案手法が最も高い結果を示している．3 つの設定と 2 つの分類モデルの計 6 つの分類精度を平均すると munit が 0.918，GP-UNIT が 0.929，提案手法が 0.942 で既存手法を大きく改善している．4.3 の定性結果や 4.4 の定量結果より，未知クラスの数を増やした場合に MUNIT では生成画像が大きく乱れたり GP-UNIT では精度が大きく低下するが，提案手法ではすべての設定で質の高い画像を生成し，精度を 0.9 以上に保つことができている．

表 1　未知クラス変化させたときの InceptionV3 を用いた画像分類の結果

	0	1	2	3	4	5	6	7	8	9	未知	既知	平均
	未知クラス：9												
MUNIT	0.985	0.988	0.958	0.953	0.944	0.941	0.969	0.938	0.947	0.848	0.848	0.958	0.947
GP-UNIT	0.995	0.993	0.971	0.981	0.974	0.967	0.982	0.977	0.980	0.952	0.952	0.980	0.977
提案手法	0.991	0.994	0.957	0.976	0.972	0.944	0.984	0.972	0.980	0.939	0.939	0.974	0.971
	未知クラス：8,9												
MUNIT	0.989	0.992	0.970	0.919	0.848	0.961	0.981	0.963	0.890	0.729	0.810	0.952	0.924
GP-UNIT	0.985	0.990	0.970	0.828	0.910	0.943	0.976	0.882	0.698	0.698	0.698	0.935	0.888
提案手法	0.973	0.990	0.954	0.941	0.828	0.954	0.969	0.946	0.888	0.668	0.778	0.950	0.911
	未知クラス：7,8,9												
MUNIT	0.969	0.939	0.914	0.874	0.836	0.937	0.972	0.844	0.794	0.715	0.784	0.920	0.879
GP-UNIT	0.976	0.984	0.880	0.924	0.924	0.852	0.976	0.762	0.861	0.866	0.829	0.930	0.900
提案手法	0.993	0.955	0.963	0.953	0.947	0.955	0.977	0.923	0.955	0.928	0.935	0.963	0.955

表 2　未知クラス変化させたときの ResNet50 を用いた画像分類の結果

	0	1	2	3	4	5	6	7	8	9	未知	既知	平均
	未知クラス：9												
MUNIT	0.985	0.989	0.956	0.956	0.965	0.954	0.970	0.968	0.976	0.920	0.920	0.968	0.963
GP-UNIT	0.994	0.983	0.972	0.991	0.948	0.968	0.985	0.980	0.993	0.924	0.924	0.979	0.973
提案手法	0.991	0.992	0.975	0.981	0.949	0.966	0.984	0.969	0.967	0.893	0.893	0.974	0.967
	未知クラス：8,9												
MUNIT	0.981	0.987	0.948	0.880	0.793	0.936	0.976	0.967	0.797	0.634	0.716	0.933	0.890
GP-UNIT	0.993	0.989	0.945	0.941	0.913	0.951	0.980	0.922	0.914	0.773	0.844	0.954	0.932
提案手法	0.972	0.989	0.941	0.935	0.905	0.945	0.971	0.943	0.888	0.792	0.840	0.950	0.928
	未知クラス：7,8,9												
MUNIT	0.984	0.926	0.950	0.857	0.921	0.921	0.978	0.813	0.895	0.797	0.835	0.934	0.904
GP-UNIT	0.988	0.966	0.928	0.850	0.894	0.919	0.893	0.893	0.857	0.776	0.842	0.932	0.905
提案手法	0.992	0.963	0.966	0.927	0.829	0.928	0.976	0.900	0.925	0.767	0.864	0.940	0.917

5 ま と め

本稿では，画像分類タスクにおいて学習データがないクラスが存在する場合に，Image-to-Image Translation を用いて学習データがないクラスの画像を生成する手法を提案した．Image-to-Image Translation において，複数のクラスの画像が学習データとして与えられた場合，形状が単純でないクラスの画像を生成して画像分類に利用すると性能が低下する問題を改善した．今後は写真などの背景にもオブジェクトが存在する場合や，クラスによって形状やオブジェクトのサイズが大きく異なる場合など未知クラスがさらに複雑な場合にも性能の低下が低減できる手法を検討したいと考えている．

文　献

[1] C. H. Lampert, H. Nickisch, S. Harmeling, ”Learning to detect unseen object classes by betweenclass attribute transfer” CVPR 2009

[2] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, ”Evaluation of output embeddings for fine grained image classification” CVPR 2015

[3] K. C. Peng, Z. Wu, J. Ernst, ”Zero-Shot Deep Domain Adaptation” ECCV 2018

[4] J. Wang, J. Jiang, ”Conditional Coupled Generative Adversarial Networks for Zero-Shot Domain Adaptation” ICCV 2019

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ”Generative adversarial nets” NIPS 2014

[6] J. Y. Zhu, T. Park, P. Isola, A. A. Efros, ”Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks” ICCV 2017

[7] M. Y. Liu, O. Tuzel, ”Coupled Generative Adversarial Networks” NIPS 2016

[8] X. Huang, Ming-Yu Liu, S. Belongie, J. Kautz, ”Multimodal Unsupervised Image-to-Image Translation” ECCV 2018

[9] Y. Shuai and J. Liming and L. Ziwei and L. C. Change, ”Unsupervised Image-to-Image Translation with Generative Prior” CVPR 2022

[10] A. Brock, J. Donahue, K. Simonyan, ”Large Scale GAN Training for High Fidelity Natural Image Synthesis” ICLR 2019

[11] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, ”Gradient-based learning applied to document recognition” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, ”Domain-Adversarial Training of Neural Networks” Journal of Machine Learning Research 17, 59 (2016), 1–35, 2016

[13] H. Xiao, K. Rasul, R. Vollgraf, ”Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms” arXiv:1708.07747 2017

[14] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, ”Contour Detection and Hierarchical Image Segmentation” IEEE TPAMI, Vol. 33, No. 5, pp. 898-916, May 2011.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, ”Rethinking the inception architecture for computer vision” arXiv:1512.00567, 2015.

[16] K. He, X. Zhang, S. Ren, J. Sun, ”Deep Residual Learning for Image Recognition” CVPR 2016

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ”ImageNet Large Scale Visual Recognition Challenge” IJCV 2015