A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions

https://dl.acm.org/doi/10.1145/3460427

Motivation

- なるべく多くのデータを利用したい（精度に直結する）
- データ収集型の限界
  - プライバシー保護の観点
  - 膨大なエッジデバイスによる通信トラフィックの増大
- 分散している学習データセットを分散させたままモデルを学習したい！！

=> Federated Learning（連合学習）

Federated Learning

- Federated Learning
  - 豊富な計算資源を持つ1台のサーバと
  - データ収集能力および，そこそこの計算資源を持つ多数のクライアントが
  - 比較的狭い帯域で通信しながら
  - 分散している学習データを分散させたまま学習する
- クライアントは個人や組織だったり，スマホや IoTデバイスだったり
- 多くの場合，深層学習を前提
- あまりにもいい資料があったので使わせてください
  - 連合学習 (Federated Learning)

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. 2017. Communication-efficient learning of deepnetworks from decentralized data. InArtificial Intelligence and Statistics. 1273-1282.

Federated Learning

- Cifer10という画像分類タスクにおいて，データを集約した場合と同程度の精度を達成できる
- Supplementary PDF of McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.

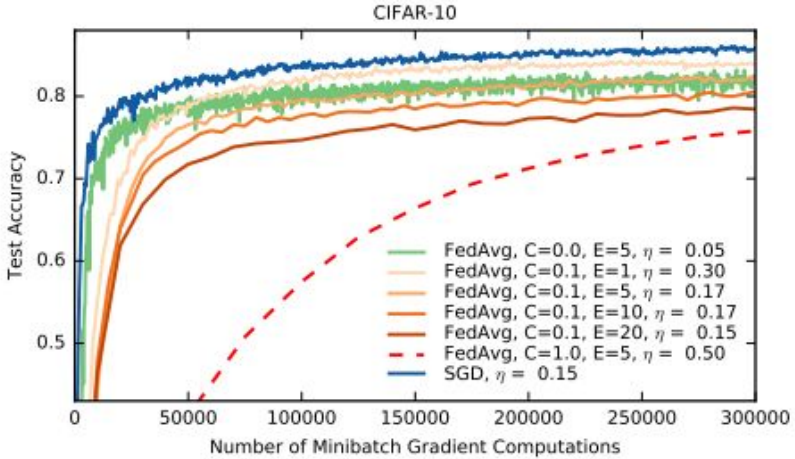


Figure 9: Test accuracy versus number of minibatch gradient computations ($B = 50$). The baseline is standard sequential SGD, as compared to FedAvg with different client fractions C (recall $C = 0$ means one client per round), and different numbers of local epochs E.

Federated Learning ができること

- データ集約型と遜色ない精度が達成できる
- 各クライアントは所有するデータ（の詳細）を公開しないでよいので，プライバシーが守られやすい
- 各クライアントが所有するデータだけで学習したモデルよりも高精度
  - 他クライアントの所有する多様で膨大なデータを利用できる
  - クライアントは参加するインセンティブがあり，途中での参加・退出も可能
- 学習データよりもモデルパラメータの方が小さいので，通信が比較的マシ
- さらに，各クライアントで特化した派生モデルを作ることも可能 (Personalization)

Federated Learning の応用例

Figure 1: Emoji predictions in Gboard. Based on the context “This party is lit”, Gboard predicts both emoji and words.

- AndroidのGboard（Googleキーボード）での文脈からの絵文字予測変換
  - 単語レベルのrecurrent neural network
  - プライバシーを確保しながら、よりスマートなモデル、低レイテンシー、低消費電力を実現
  - 共有モデルのアップデートに加えて、改善されたモデルをすぐに使える
  - Ramaswamy, Swaroop, et al. "Federated learning for emoji prediction in a mobile keyboard." arXiv preprint arXiv:1906.04329 (2019).
  - Google AI Blog: Federated Learning: Collaborative Machine Learning without Centralized Training Data
- 医療応用（個々の患者の情報は出さずに珍しい疾患に対しても適切な処置につながるかも？）
  - https://blogs.nvidia.co.jp/2020/01/08/what-is-federated-learning/

Distributed Machine Learning vs. Federated Learning

Distributed Machine Learning

- 学習データを一度集約して分配
- 学習処理の分散・並列・高速化が目的
- 計算ノードとして計算タスクが割り振られる

Federated Learning

- 学習データはクライアントごとに異なる
  - それぞれ異なる分布を持つ

Split Learning vs. Federated Learning

Split Learning

- 各クライアントはカットレイヤーと呼ばれる特定の層までのモデルを学習する
- 残りの層は全クライアント共通で，サーバ側で学習される
- カットレイヤーのやり取りのみ通信

Federated Learning

- 各クライアントおよびサーバ両方がモデル全体を持つ
- モデルパラメータまたは勾配の通信を伴う

Federated Learning の課題

- 学習の収束が遅い
- クライアントのデータの分布がそれぞれ異なる
- モデルの共有に大きな通信トラフィックが繰り返し発生
- 悪意のあるクライアント等が学習の邪魔をする場合
- セキュリティの問題
    - プライバシーが必ず守られるわけでない

Federated Learning の課題

- Non-iid(independent and identically distributed)
  - 各Clientの観測データの従う分布が全 Clientのデータを集約した場合の分布と一致しない
  - 全体としてモデルの更新がうまくいかない
- SCAFFOLD
  - グローバルモデルの更新方向を各クライアントで予測し，ローカルモデルの更新方向を調整
  - S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. InProceedings of the International Conference on Machine Learning, Vol. 119.5132-5143.
- FedAdp
  - 学習への貢献度をもとに，各クライアントの重みを調整
  - H.Wu and P.Wang, “Fast-Convergent Federated Learning with Adaptive Weighting” IEEE transactions on Cognitive Communications and Networking, vol. 7, no. 4, pp. 1078-1088, Dec. 2021

Federated Learning の課題

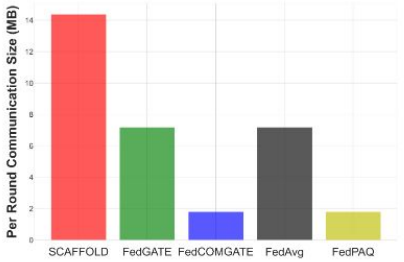


Figure 3: Communication cost at each round for the CIFAR10 dataset with a 2-layer MLP.

- モデルの共有に大きな通信トラフィックが繰り返し発生
- FedCOM
  - パラメタの量子化，枝刈り，などでモデルを圧縮し，通信を削減
  - F. Haddadpour, et al., “Federated Learning with Compression: Unified Analysis and Sharp Guarantees.” AISTATS 2021.
- 空中計算
  - 電波の重畳をモデルの統合に利用し多数のクライアントがいる状況でのトラフィックの負荷を軽減？
  - K. Yang, T. Jiang, Y. Shi and Z. Ding, "Federated Learning via Over-the-Air Computation," in IEEE Transactions on Wireless Communications, vol. 19, no. 3, pp. 2022-2035, March 2020, doi: 10.1109/TWC.2019.2961673.

Federated Learning の課題

- モデルの共有に大きな通信トラフィックが繰り返し発生
- Fed Fog
  - 計算資源と通信機能を備えた Fog Serverをクライアントの近くに配置
  - Fogを介して入れ子の構造をつくることで，通信負荷を軽減
  - Nguyen, Van-Dinh, et al. "FedFog: Network-Aware Optimization of Federated Learning over Wireless Fog-Cloud Systems." *arXiv preprint arXiv:2107.02755* (2021).

Federated Learning の課題

- 悪意のあるクライアント等が学習の邪魔をする場合
- Byzantine-Robust Distributed Learning
  - Byzantine fault tolerance
    - 分散コンピューティングにおいて、個々のオブジェクトにおける故障や故意，または通信によって偽の情報を伝達する可能性がある場合であっても，全体として正しく動作すること
  - サーバーが各クライアントの送信モデルを集約する際に異常値を除いてやる
  - Dong Yin, Yudong Chen, Ramchandran Kannan, Peter Bartlett, “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates.” Proceedings of the 35th International Conference on Machine Learning, PMLR 80:5650-5659, 2018.

Privacy-preserving Machine Learning: Motivation

- 世界的なプライバシー保護に関する法規制
- プライバシーとは、「個人が、自分に関するどのような情報が収集・保存され、その情報が誰によって、誰に対して開示されるかを制御し、または影響を及ぼすことを保証すること」
  - W. Stallings. 2017.Cryptography and Network Security Principles and Practices(7th ed.). Pearson Education, Inc.
- Privacy-preserving Machine Learning
  - 大まかに3つの一般的なプライバシー保護機構
    - Cryptographic Techniques
    - Perturbative Techniques
    - Anonymization Techniques

Cryptographic Techniques

- Homomorphic encryption
  - データの特徴を保持する暗号化によって，暗号化されたまま一部の計算が可能な暗号化方式
  - 暗号化方式が演算★についてHomomorphicであるとは，以下が成り立つことをいう

$$E(m_1) \star E(m_2) = E(m_1 \star m_2), \forall\, m_1, m_2 \in M$$

  - ただし，Eは平文を受け取って暗号を返す関数，Mはあり得るすべての平文の集合
- Partially homomorphic encryption
  - 加算(XOR)と乗算(AND)など1種類の演算を何度でも実行できる
- Fully homomorphic encryption
  - 任意の演算を何度でも実行できる
  - 任意のブール関数は加算(XOR)と乗算(AND)，定数で表現できるため，加算(XOR)と乗算(AND)についてhomomorphic であればよい
  - 強力だが計算コストが高い

Cryptographic Techniques

- Secret sharing A. Shamir. 1979. How to share a secret.Commun. ACM22, 11 (1979), 612-613.
  - 一人のディーラーが参加者 n人に一つずつ秘密鍵 viを配る
  - (t,n)-threshold secret sharing
    - t個以上の秘密鍵で秘密 sを復号できる
    - t個未満の秘密鍵で秘密 sを復号できない
  - ディーラーの不正や悪意のある参加者集団には脆弱
  - t == n の場合 s = v_1 XOR v_2 XOR … XOR v_n とすればよい

Cryptographic Techniques

- Shamir's Secret Sharing
  - 定数項が秘密s，残りの係数がランダムである (t-1)-次多項式f(x)を決める
  - 参加者一人一人に秘密鍵 v_i = (x_i, f(x_i)) を配る
  - t個の点を通る t-1次多項式f(x)は一意に定まる
  - このとき s = f(0) となる
  - 秘密鍵の一部が漏洩した場合
    - 定数項が0，残りの係数がランダムである (t-1)-次多項式 f'(x)を決める
    - 参加者一人一人に秘密鍵（の更新？） v'_i = (x_i, f'(x_i)) を配る
    - 各参加者は秘密鍵 v_i ← v_i + v'_i で更新する
    - t個の点を通る t-1次多項式 f(x)+f'(x) は一意に定まる
    - このとき s = f(0)+f'(0)
    - 秘密sを変更せずに鍵の新調が可能！！

Cryptographic Techniques

- Secure Multi-party Computation (SMC) Protocol
  - Multi-party Computation (MPC) と表記されるほうが多い
  - n人の参加者{P1,P2,...,Pn}が各自のデータセットDiを持つ
  - Diを明かすことなく共同で目的関数 f(D1,D2,...,Dn) を計算する
  - 信頼できる第三者を必要としない
  - https://www.jaist.ac.jp/~fujisaki/2018/lec-mpc-kanazawa-20180629.pdf

Perturbation Techniques

- 元データに意図的なノイズを加えることで、加工済データから計算される統計情報は元データと統計的に区別がつかなくする
- もちろん加工しない場合よりは劣化する
- 単純で効率的，データ分布に関する知識が不要
- https://www.jstage.jst.go.jp/article/jssst/29/4/29_4_40/_pdf
- https://www.slideshare.net/kentarominami39/ss-64088396

Anonymization Techniques

- 公開されたデータの有用性を維持しながら，識別可能な情報を削除する
  - 識別子とは単独で個人を特定できる属性，匿名化によって隠される
  - 準識別子は単独で個人を特定できない属性，比較的容易にアクセスできる
  - しかし，このような条件下でも準識別子の組合せから個人を特定することが可能
  - k-anonymity
    - 任意の準識別子の組合せが全く同一の個人が少なくとも k人以上存在するという要件
    - ある人のデータをデータベース中から k 個未満に絞り込めない
  - l-diversity
  - t-closeness

Anonymization Techniques

- Differential privacy
  - データセット中のインスタンスの個人情報の開示の程度を定量化する指標
  - “高々1要素だけ異なるデータベースからある値が出力される確率の対数の距離が εで抑えられる”
  - 気持ちとしては多分，“任意の個人の情報が含まれていていなくても出力の差分が十分小さい”

定義 2.1. $\mathscr{D}$ をデータベースの集合，d を非負整数とする．プライバシー機構 $\mathcal{K} : \mathscr{D} \to \mathbb{R}^d$ を確率的アルゴリズム，ε を (小さな) 正実数とする．$\mathcal{K}$ が ε-*Differential Privacy* を与えるとは，任意の $S \subseteq \mathrm{Range}(\mathcal{K})$ と，“高々1 要素だけ異なる” ような任意のデータベースのペア D_1, D_2 に対して以下が成立することである．

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\varepsilon) \Pr[\mathcal{K}(D_2) \in S] \quad (1)$$

ただし $\mathrm{Range}(\mathcal{K})$ は $\mathcal{K}$ の値域，すなわち $\mathcal{K}$ が出力する可能性のある値の集合である．

$$|\log \Pr[\mathcal{K}(D_1) \in S] - \log \Pr[\mathcal{K}(D_2) \in S]| \leq \varepsilon \quad (2)$$

https://www.jstage.jst.go.jp/article/jssst/29/4/29_4_40/_pdf より

Privacy-preserving Metrics

- データセットのプライバシー損失を測定するプライバシー指標
  - 一般的に、プライバシー指標はケースバイケース
  - 単一のメトリックではプライバシーを完全に評価することができない
- 保護されたデータのUtilityを測定するUtility指標
  - 一般的に，元のデータと保護されたデータの間の類似性を測定
  - 保護されたデータが元データの統計情報をどの程度保持しているか
  - 保護されたデータと元のデータを使用したタスクの評価精度

Privacy-preserving Federated Learning

cenario 1: Who might be a malicious adversary?

ー ー ー

- 内部アクター（クライアントとサーバ）/外部アクター（消費者と盗聴者）
  - 消費者は最終モデルにアクセスできる
  - 盗聴者は参加者とサーバ間の通信を傍受し，途中の更新や最終モデルを盗み見る
  - 悪意を持ったクライアントは
    - 何かしらの意図的に設計された学習結果をサーバに送信する可能性がある
    - 学習途中の更新やモデルから他のクライアントのデータセットを探ることができる
  - 悪意を持ったサーバは
    - 中間更新とモデルからクライアントのデータセットを探ることができる
    - 特定クライアントに意図的に設計されたモデルを共有する，または参加するクライアントを選択・規制することで，特定クライアントのデータセットを探ることができる
  - 悪意を持ったモデルの消費者は
    - 最終モデルからクライアントのデータセットを探ることができる
  - 悪意を持った盗聴者は
    - 学習途中の更新やモデルから他のクライアントのデータセットを探ることができる

Scenario 2: What types of privacy attacks?Passive and active attacks

---

- 受動的攻撃
  - 正規のプロトコルからは逸脱しないが、プロトコルから得た情報から参加者の秘密情報を得ようとする
  - 受動的ブラックボックス攻撃
    - 敵対者はクエリ結果のみからいろいろ推定
  - 受動的ホワイトボックス攻撃
    - 敵対者は中間学習更新、モデルパラメータ、およびクエリ結果にアクセス
- 能動的攻撃
  - 正規のプロトコルから逸脱し，参加者の秘密情報を得ようとする
  - 例えば何かしらの意図のある勾配をアップロードする等
    - 学習を阻害する
    - 他クライアントの情報を推定する

cenario 3: When might a data privacy leakage occur?

- 学習フェーズにおけるプライバシー漏洩のリスク
  - 学習フェーズにおけるすべての更新情報が、悪意のある敵にさらされる可能性がある
  - 局所勾配、局所モデル重み、集約された勾配やモデル重み、そして最終的なモデル
  - 更新や最終モデルからのプライバシー漏洩を制限するために、 FL学習プロセスにおいて差分プライバシーを適用することができる
- 参照フェーズでは、プライバシー漏洩のリスク
  - (1)モデルパラメータに基づく攻撃
  - (2)モデルクエリに基づく攻撃

cenario 4: Where might a data privacy leakage occur? Weight update, gradient update, andthe final model.

- 勾配更新をやり取りするFLフレームワーク
- 重み更新をやり取りするFLフレームワーク
- 最終的なモデル

cenario 5: Why might a malicious attacker launch an attack?

- inference of class representatives
  - 学習データセットに含まれる実際のデータではなく，合成された一般的なサンプルであるクラス代表を抽出
  - 例えば，顔認識モデルなどから人種は推定できる
- inference of a membership
  - あるデータサンプルが学習用データセットに含まれているかどうかを判断することを目的と
  - 例えば、ある病気に関する分類器を学習するのに特定の患者の記録が使用されたかどうか
- inference of the properties of training data
- inference of training samples and labels

具体例

- Differential privacyにより保護されたモデルであっても，Generative Adversarial Network（GAN）で非公開データが再現可
  - B. Hitaj, G. Ateniese, and F. Perez-Cruz. 2017. Deep Models under the GAN: Information leakage from collaborativedeep learning. InProceedings of the ACM SIGSAC Conference on Computer and Communications Security. 603-618.

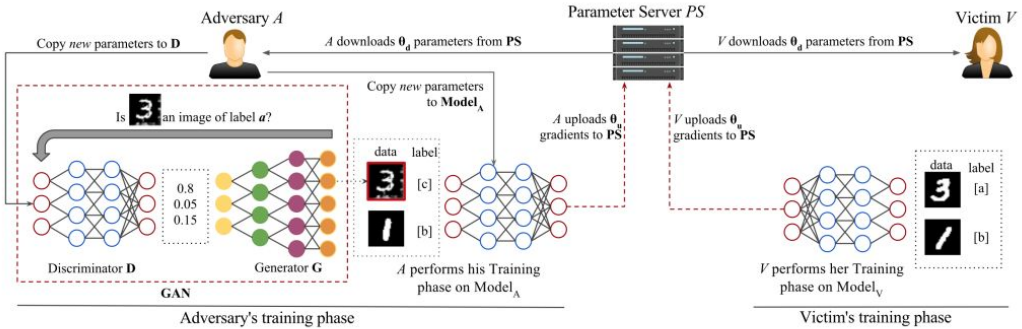


Figure 4: GAN Attack on collaborative deep learning. The victim on the right trains the model with images of 3s (class *a*) and images of 1s (class *b*). The adversary only has images of class *b* (1s) and uses its label *c* and a GAN to fool the victim into releasing information about class *a*. The attack can be easily generalized to several classes and users. The adversary does not even need to start with any true samples.

Figure 2: Picture of Alice on the victim's phone, X, and its GAN reconstruction, X'. Note that $X' \neq X$, and X' was not in the training set. But X' is essentially indistinguishable from X.

Figure 3: GAN-generated samples for the 'horse' class from the CIFAR-10 dataset

具体例

- 悪意のあるサーバは共有するモデル，他クライアントも含めた更新情報などから特定クライアントの非公開データを再現可
  - Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi. 2019. Beyond inferring class representatives: User-level privacy leakage from federated learning. InProceedings of the IEEE Conference on Computer Communications.2512-2520

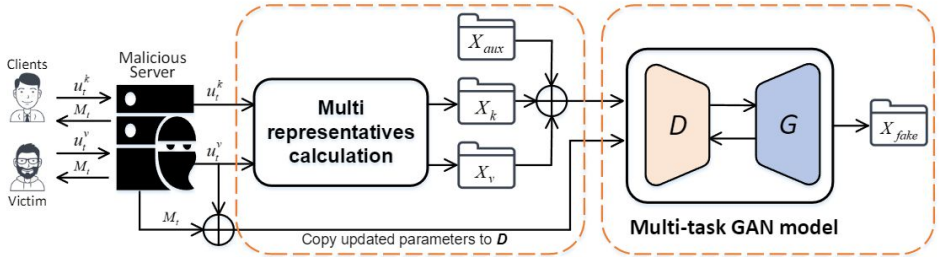
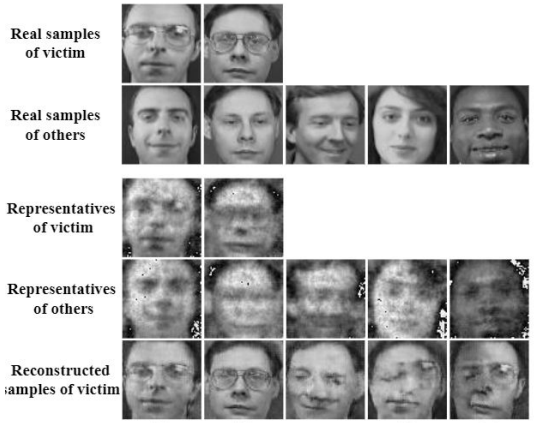


Fig. 2: Illustration of the proposed mGAN-AI from a malicious server in the federated learning. There are N clients, and the vth client is attacked as the victim. The shared model at the tth iteration is denoted as M_t, and u_t^k denotes corresponding update from the kth client. On the malicious server, a discriminator D (orange) and generator G (blue) are trained based on the update u_t^v from the victim, the shared model M_t, and representatives X_k, X_v from each client. X_{aux} denotes an auxiliary real dataset to train D on the real-fake task.



具体例

- 公開される勾配のみから非公開データを再現できる
  - L. Zhu, Z. Liu, and S. Han. 2019. Deep leakage from gradients. InAdvances in Neural Information Processing Systems,Vol. 32. 14774-14784.

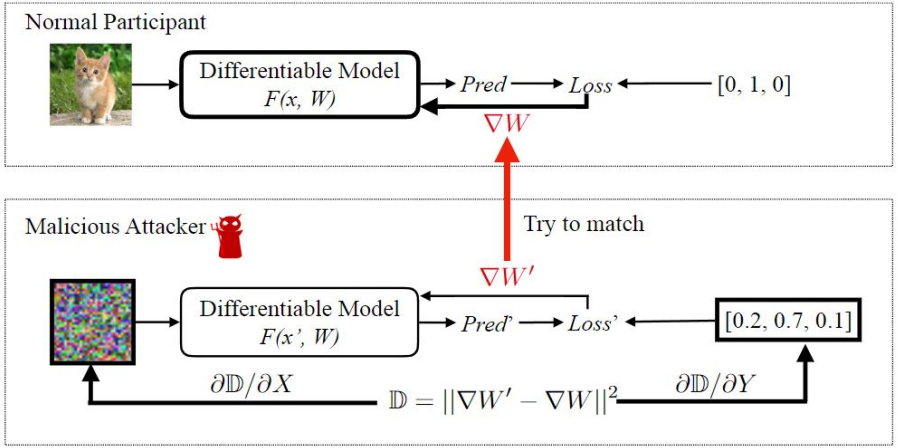


Figure 2: The overview of our DLG algorithm. Variables to be updated are marked with a bold border. While normal participants calculate ∇W to update parameter using its private training data, the malicious attacker updates its dummy inputs and labels to minimize the gradients distance. When the optimization finishes, the evil user is able to obtain the training set from honest participants.

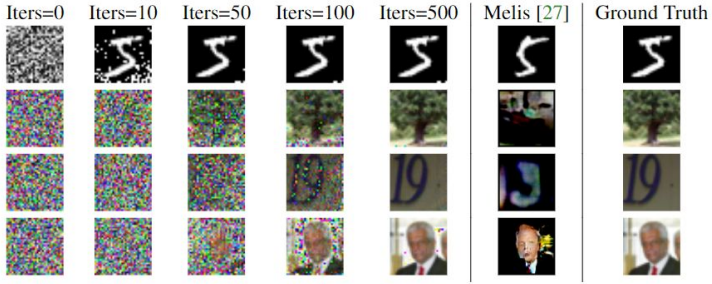


Figure 3: The visualization showing the deep leakage on images from MNIST [22], CIFAR-100 [21], SVHN [28] and LFW [14] respectively. Our algorithm fully recovers the four images while previous work only succeeds on simple images with clean backgrounds.

Summary

- FLの急速な発展にもかかわらずこの研究分野は依然として課題が山積