



4KリアルタイムAI推論を実現する超高精細映像AI推論LSIと道路監視への応用

4K映像の画素情報を活かしたまま、リアルタイムかつ低電力でAI（人工知能）推論できる超高精細映像AI推論LSI（Large Scale Integration：大規模集積回路）を開発しました。本LSIでは、4K映像を高精細に推論する技術と、演算量・データ転送量を削減する処理効率化技術により、高精細推論とリアルタイム・低電力処理を両立しています。本稿では、これらの技術と評価結果に加え、モビリティ分野への応用例として、4Kの高精細さを活かして遠方の小さな危険事象もとらえる道路監視への適用について紹介します。

キーワード：#高精細、#推論、#生成AI

よしだ	しゅうへい	はぎわら	なるき
吉田	周平	/萩原	成基
えばら	なるのり	はやみ	たいが
榎原	成則	/速見	泰雅
いぬま	ゆうこう	おおもり	ゆうや
飯沼	宥光	/大森	優也
ほりした	ゆうすけ	うざわ	ひろゆき
堀下	祐輔	/鶴澤	寛之

NTTデバイステクノロジーセンタ

はじめに

深層学習技術の進展に伴い、カメラ映像をAI（人工知能）で解析し、実世界の状況を把握する映像AI推論の活用が広がっています。特に、道路監視やドローン、人流分析などのエッジ・端末アプリケーションでは、高精細カメラ1台で広い範囲をとらえ、物体をリアルタイムかつ低電力で検出することが求められています。

このような電力制約の厳しいエッジ・端末では、サーバ等で利用される数100 W級のGPUに対し、消費電力が1桁以上低い数10 W級のAIデバイスが利用されています。一方、一般にAI推論モデルでは、計算量の低減や学習容易性の観点から入力画像サイズに制約があります。例えば物体検出のAI推論モデルであるYOLO（You Only Look Once）v3⁽¹⁾の公式モデルの最大サイズは608×608ピクセルです。近年では、より大きな画像を扱えるモデルも登

場していますが、4K（3840×2160ピクセル）と比べると依然として低解像度です。そのため、4Kカメラで撮影してもAI推論時には映像を縮小することが一般的であり、遠方に小さく映る物体の画素情報が失われ、検出が困難になるという課題があります。

NTTでは、この課題を解決するため、高精細映像AI推論ハードウェアの研究開発に取り組み、4K映像を高精細なままリアルタイムかつ低電力でAI推論可能な超高精細映像AI推論LSI（Large Scale Integration：大規模集積回路）を開発しました^{(2),(3)}（図1）。本LSIでは、一般的なエッジ・端末向けAIデバイスで608×608ピクセルに縮小してYOLOv3を実行した場合と同等以下の消費電力（20 W以下）で、4K解像度のリアルタイム物体検出（30 fps）を実現しています。

本稿では、4Kリアルタイム推論を実現する主要技術と、4K映像の高精細さを活かした道路監視への適

用について紹介します。

超高精細映像AI推論LSI

開発した超高精細映像AI推論LSIの構成を図2に示します。本LSIは、AI推論エンジン、画像処理エンジン、映像符号化エンジン、CPU、外部メモリーインタフェースなどを集積し、4K映像に対する画像処理、AI推論、映像符号化を1チップで実行できます。

4K映像では処理すべき画素数が大幅に増えるため、AI推論に必要な演算量やデータ転送量も増大します。一方、エッジ・端末では電力やハードウェア資源に制約があるため、単に演算器の性能を高めるだけでは、4K映像のリアルタイム・低電力AI推論を実現することは困難です。

そこで本LSIでは、4K映像の画素情報を活かすAI推論高精細化技術^{(4),(5)}に加え、「できるだけ計算しない」「できるだけデータを動かさない」ことを基本思想とした処理効

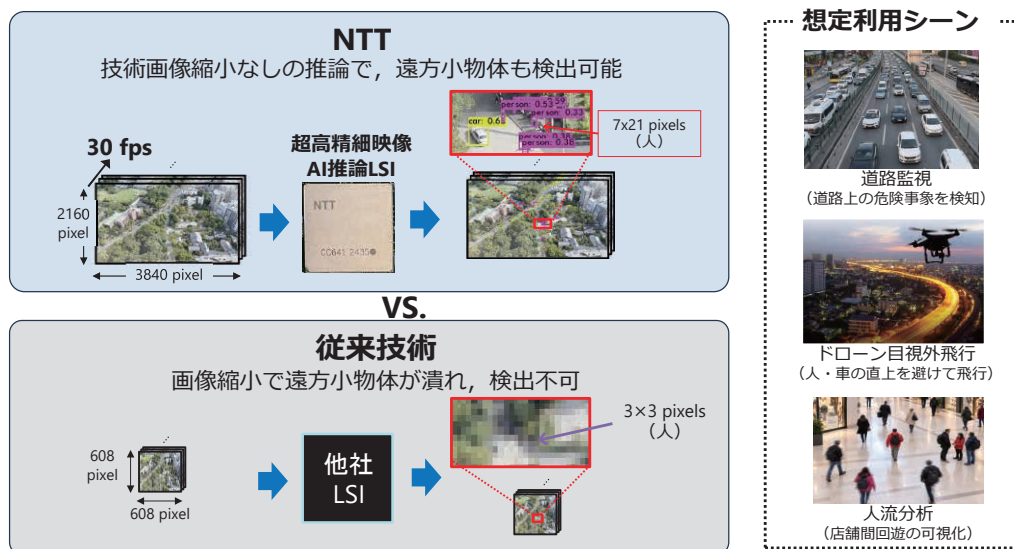


図1 超高精細映像 AI 推論 LSI の特徴と想定利用シーン

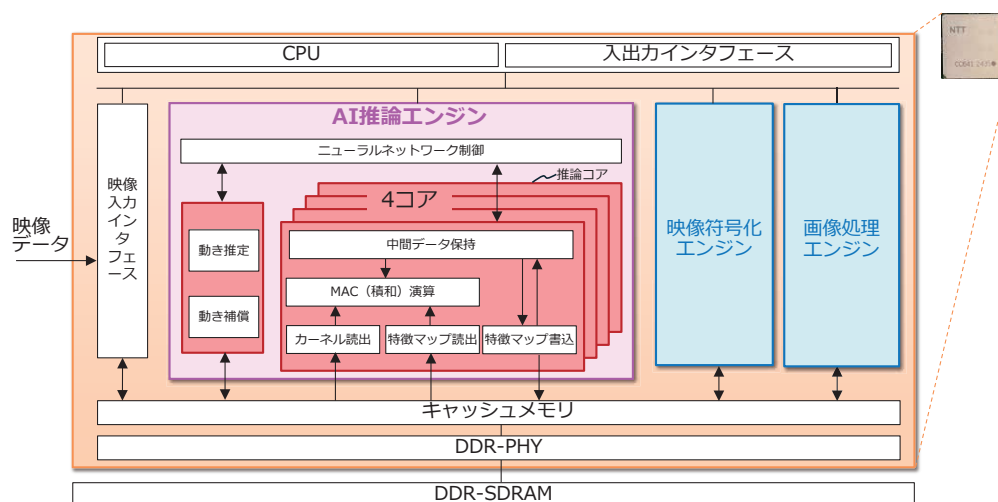


図2 超高精細映像 AI 推論 LSI の構成

率化技術を導入しています。具体的には、映像の時間方向の相関を利用して AI 推論の演算量を削減するとともに、層統合や適応型マルチコア並列化によって外部メモリとのデータ転送量を削減します。これらの技術を組み合わせることで、4K 映像のリアルタイム・低電力 AI 推論を実現しています。以下では、これら

の主要技術について説明します。

■ AI 推論高精細化技術

本技術の処理イメージを図 3 (a) に示します。4K 映像を縮小することなく推論するため、4K 映像を AI モデルの入力画像サイズに応じて複数の領域に分割し、分割画像ごとに推論を行います。これにより、4K 映像が持つ画素情報を維持し、従来の

縮小処理では検出が難しかった小さな物体を検出できます。

一方、分割領域をまたぐ大きな物体を検出するため、4K 映像全体を縮小した画像に対する推論も行います。全体画像と分割画像の推論結果を合成することで、大きな物体の検出性能を維持しながら遠方に小さく映る物体まで検出し、4K 映像の高

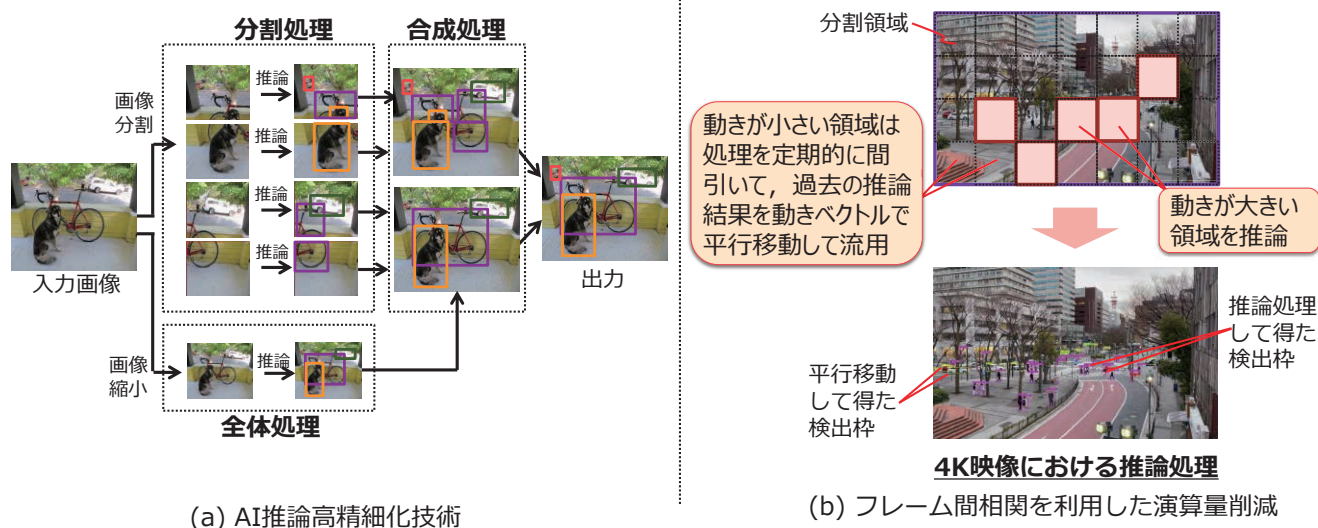


図3 AI推論高精細化技術とフレーム間相関を活用した推論演算量削減技術

い空間解像度を活かした物体検出を実現します。

■フレーム間相関を活用した推論演算量削減技術

AI推論高精細化技術では4K映像を多数の領域に分割して推論するため、そのままでは演算量が膨大となります。そこで本LSIでは、動画のフレーム間相関を活用し、すべてのフレーム、すべての領域に対してAI推論を行うことを回避することで、演算量を削減します（図3(b)）。

動画では連続するフレーム間で映像内容の変化が小さく、高い相関があります。本LSIでは、AI推論エンジン内の動き推定・補償部で映像内の動きを推定・補償し、その結果をAI推論に活用します。動きが小さい領域ではAI推論を定期的に間引き、過去フレームの推論結果を動きベクトルに基づいて位置補正して再利用する一方、変化の大きい領域ではAI

推論を実行します。これにより、すべての分割領域を毎フレーム一律に推論することを避け、推論性能を維持しながら「できるだけ計算しない」処理を実現します。

■層統合・適応型マルチコア並列化によるデータ転送量削減技術

AI推論を低電力で実行するには、演算量に加えて外部メモリとのデータ転送量を削減することが重要です。CNN（Convolutional Neural Network）では各層で生成された中間データ（特徴マップ）を次の層へ受け渡します。このデータを、その都度外部メモリへ書き出して読み戻すと、大量のデータ転送が発生します。

本LSIでは、特徴マップを小さな領域（タイル）に分割し、1つのタイルについて複数層を連続処理する層統合を導入しています（図4(a)）。層間の中間データをLSI内部に保持

することで中間層の外部メモリアクセスを省略でき、タイル単位で保持することで必要な内部メモリ容量も抑えられます。

また、本LSIでは4つの推論演算コアを用いて並列処理します。CNNの前半層では特徴マップが大きく、後半層ではカーネルのデータ量が大きくなるため、前半層では特徴マップを領域ごとに、後半層ではカーネルを分割して各コアに割り当てます。層ごとにデータ量の大きい側を分割することで重複読出しを抑え、外部メモリとのデータ転送量を削減します（図4(b)）。これらにより、「できるだけデータを動かさない」処理を実現しています。

性能評価

まず、層統合と適応型マルチコア並列化による外部メモリとのデータ

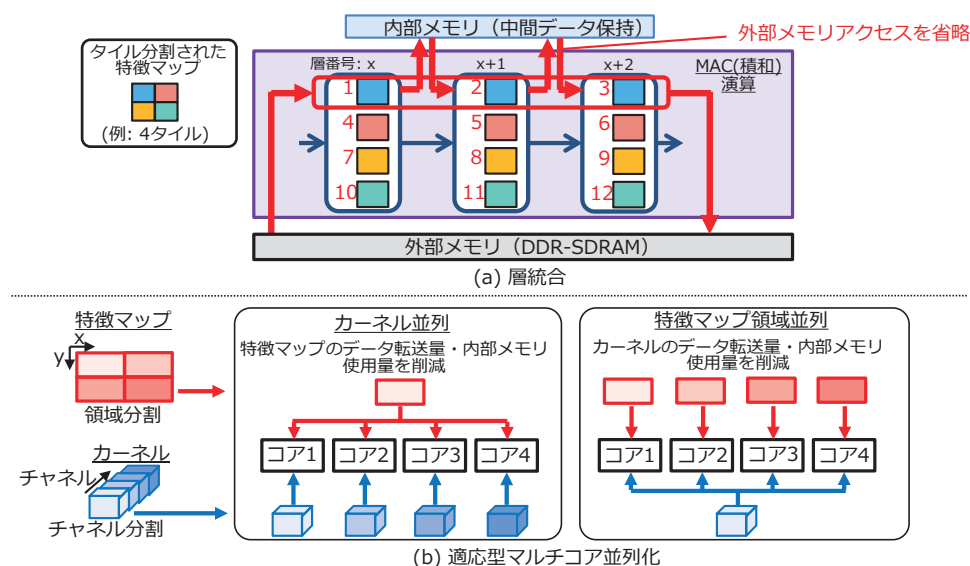


図4 層統合と適応型マルチコア並列化によるデータ転送量削減技術

転送量の削減効果を評価しました。YOLOv3, ResNet50, VGG16 を用いて評価した結果、これらの技術を適用しない場合と比較して、外部メモリとのデータ転送量を53～78%削減できることを確認しました(図5(a))。

次に、4K映像に対する物体検出の推論精度とスループットを評価しました。YOLOv3と3種類の4K映像を用い、①4K映像全体を608×608ピクセルに縮小して推論する従来手法、②AI推論高精細化技術だけを適用する手法、③AI推論高精細化技術とフレーム間相関による演算量削減技術を組み合わせた提案手法を比較しました。

図5(b)に示すように、①では高いスループットが得られる一方、画像縮小により推論精度が低下します。②では推論精度は向上しますが、多数の分割画像を推論するためスルー

ットが大きく低下します。これに対し③では、高精細化による推論精度を維持しながら30 fpsのスループットを実現し、高い推論精度とリアルタイム処理を両立できることを確認しました。

4Kの高精細さを活かした道路監視への応用

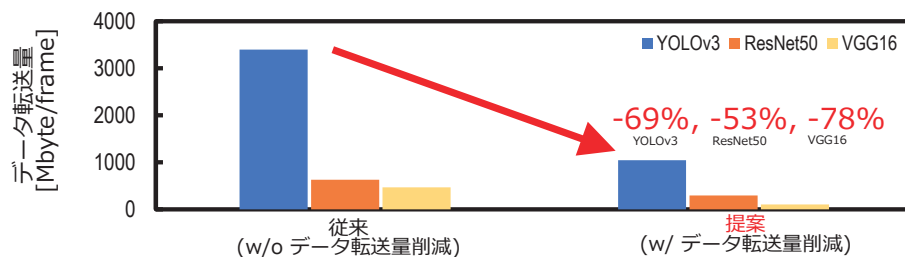
本LSIの特長を活かし、モビリティ分野への応用例として、道路監視における危険事象の遠距離検知への適用検討を進めています。

高速道路などでは、停止車両、逆走車両、落下物などの危険事象を早期に把握することが重要です。遠方の物体ほど映像中に占める画素数が少なくなるため、AI推論時に画像を縮小すると物体の特徴が失われやすく、検出が難しくなります。本LSIでは、4K映像の画素情報を活かす

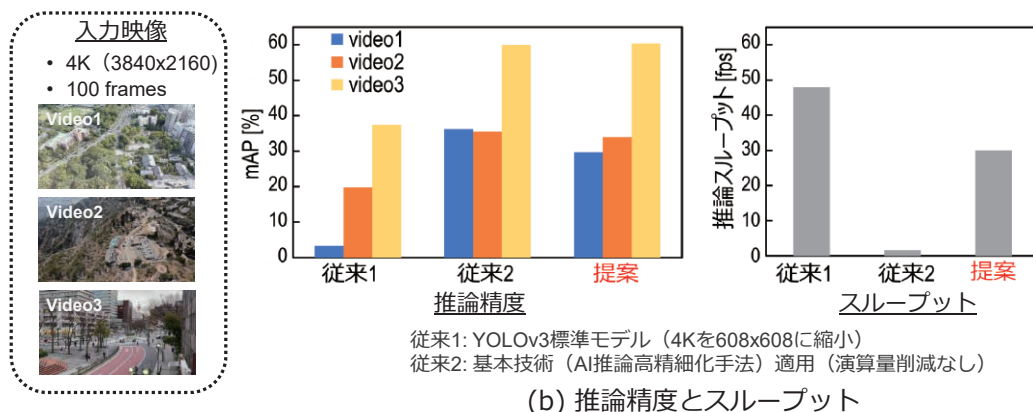
ことで、遠方に小さく映る車両や落下物などを検出します。より遠方から検知できれば、危険事象の早期発見に加え、1台のカメラで監視できる範囲を広げ、道路監視設備の効率化も期待できます。

また、本LSIによる車両の検出結果を時系列に追跡し、位置変化から移動量や移動方向を求めることで、停止車両や逆走車両を検知します。本LSIは4K映像に対して30 fpsで物体検出できるため、高速に走行する車両についても短い時間間隔で位置情報を取得でき、物体追跡や移動方向の判定に活用できます。

このように、本LSIを中核として、高精細な物体検出と時系列処理を組み合わせることで、道路上で発生する危険事象を遠方からリアルタイムにとらえることが可能になります。



(a) 外部メモリへのデータ転送量



(b) 推論精度とスループット

図 5 性能評価結果

今後の展開

本稿では、4K映像を高精細なままリアルタイムかつ低電力でAI推論する超高精細映像AI推論LSIと、その主要技術および道路監視への適用について紹介しました。

今後は、生成AIとの連携についても検討を進めていきます。車両や落下物などを高精細にとらえるだけでなく、それらが事故などによる危険な状況にあるのか、あるいは通常の交通状況の一部なのかを生成AIで判断することで、人による確認・対応が必要な事象をよりの確に把握し、監視業務の省力化や重要事象への初動の迅速化につながると考えています。こうした連携を通じて、4K

映像の高精細さを活かした映像解析と、生成AIによる状況理解を組み合わせる「AIの目」として発展させるとともに、道路監視にとどまらず、高精細映像の活用が求められるさまざまな分野への展開をめざします。

参考文献

- (1) <https://arxiv.org/abs/1804.02767>
- (2) <https://group.ntt.jp/newsrelease/2025/04/11/250411a.html>
- (3) Y. Omori et al.: "A 0.07 μ J/pixel 4K 30fps Video AI Inference Accelerator with Video Transmission Functionality for Real-Time Edge Devices," Proc. of IEEE A-SSCC 2025, pp. 1-3, 2025.
- (4) H. Uzawa et al.: "High-definition object detection technology based on AI inference scheme and its implementation," IEICE Electronics Express, Vol. 18, No. 22, p. 20210323, 2021.
- (5) H. Uzawa et al.: "High-definition technology of AI inference scheme

for object detection on edge/terminal," IEICE Electronics Express, Vol. 20, No. 13, p. 20232002, 2023.



(上段左から) 吉田 周平 / 萩原 成基 / 榎原 成則 / 速見 泰雅
 (下段左から) 飯沼 宥光 / 大森 優也 / 堀下 祐輔 / 鶴澤 寛之

「高精細な映像を、高精細なままAIで活かす」ことをめざして研究開発を進めています。本技術をさまざまな実フィールドへ展開し、映像からより多くの価値を引き出せるAIシステムの実現につなげていきます。

◆問い合わせ先

NTTデバイステクノロジーセンタ