EasyChair Preprint

№ 13681

Real-Time Unsupervised Classification in Industrial Time-Motion Studies Using 2D Key Points

Wetu Vexo, Chawalit Jeenanunta, Sapa Chanyachatchwan, Apinun Tunpan and Nisit Sirimarnkit

June 16, 2024

Real-time Unsupervised Classification in Industrial Time-Motion Studies Using 2D Key Points

*

Wetu Vexo
School of Management Technology, Sirindhorn International of Technology, Thammasat University
Phatum Thani, Thailand
m6522040622@g.siit.tu.ac.th

Chawalit Jeenanunta*
School of Management Technology, Sirindhorn International of Technology, Thammasat University
Phatum Thani, Thailand
chawalit@siit.tu.ac.th ORCID: 0000-0002-1932-9776
*Corresponding Author

Sapa Chanyachatchwan
National Electronics and Computer Technology Center, National Science and Technology Development Agency
Phatum Thani, Thailand
sapa.chanyachatchawan@nectec.or.th

Apinun Tunpan
SMART Sense Industrial Design Co., Ltd,
Bangkok, Thailand
apinun@smartsensedesign.com

Nisit Sirimarnkit
SMART Sense Industrial Design Co., Ltd,
Bangkok, Thailand
nisit@smartsensedesign.com

***Abstract*— Time-motion studies (TMS), essential for analyzing and optimizing work processes in industrial environments, have traditionally relied on manual observation and data collection, incurring labor-intensive efforts and potential biases. The advancement of computer vision and machine learning presents opportunities for automating and refining TMS through real-time 2D key point data analysis. However, existing methods often require labeled data for training, which can be a significant bottleneck. To address this challenge, we propose a novel approach leveraging unsupervised learning to classify human motions without requiring any labeling, thereby streamlining the process significantly. This study utilizes the Mediapipe framework to extract human skeletal key points, which are processed by a sequence-to-sequence (seq2seq) autoencoder model. The Encoder component captures key point sequences while the Decoder reconstructs these sequences from a compressed latent space. Subsequently, unsupervised clustering techniques are applied to group similar activities, enhancing action recognition efficiency without manual labeling. This innovative methodology eliminates the dependency on labeled data and paves the way for more efficient and scalable TMS in industrial settings.**

Keywords: Pose Classification, Pose Estimation, Unsupervised Learning, Real-time Processing, Time-Motion Studies

I. INTRODUCTION

Action recognition is essential for monitoring human workers in industrial settings to enhance operational efficiency, safety, and quality control. By accurately identifying and analyzing worker actions, industries can ensure compliance with standard operating procedures, promptly detect and address errors, and optimize workflow management. Traditional methods for monitoring workers in industrial environments, such as direct observation or video analysis, are often employed but are often time-consuming and labor-intensive.

Human Action Recognition (HAR) is a part of computer vision that focuses on how computers and machines can learn and recognize the pattern of the input data from various sources, including sensors, video sequences, or real-time video, and then predict and classify the action being performed. HAR using 2D keypoint data has emerged as a promising approach for understanding and analyzing human behavior in real time, particularly in industrial settings.

Recent research has extensively explored Deep Learning-based methods involving Convolution Neural Networks (CNNs), Autoencoder, Recurrent Neural Networks (RNNs), and other Deep Neural Architecture [1], [2], [3], [4], [5], [6].

Integrating HAR systems within industrial environments holds immense potential for enhancing worker safety, optimizing workflow management, and improving overall operational efficiency. By accurately identifying and classifying worker actions in real-time, these systems can promptly detect and rectify errors and provide invaluable insights into worker productivity and task performance. Moreover, integrating HAR with time-motion studies allows for the precise measurement and analysis of task durations, enabling data-driven decision-making for process optimization and efficiency improvement. [7], [8].

Despite the significant advancements in HAR, several challenges remain, particularly in supervised learning. While supervised methods can achieve remarkable accuracy, they depend heavily on extensive labeled datasets, which are often costly and time-consuming to create. This dependency on labeled data poses a significant obstacle to the widespread adoption of HAR in industrial environments, where the diversity of activities and environments necessitates highly customized and adaptable models. To overcome the limitations of supervised learning, exploring unsupervised learning techniques, which do not require labeled data, holds immense promise for developing more efficient, scalable, and adaptable HAR systems.

II. RELATE WORK

Real-time unsupervised classification in industrial time-motion studies using 2D key points or real-time video data is a modern research area that aims to automate the categorization of movements in industrial settings without manual labeling or supervision. Many researchers are exploring related concepts that can help develop systems capable of real-time motion data classification and analysis. A landmark study in this domain was conducted by Srivastava et al. (2015), who introduced an unsupervised learning approach for video representation using Long Short-Term Memory (LSTM) networks [6]. Their research showcased the

effectiveness of LSTMs in processing sequential images for tasks like frame reconstruction, future frame prediction, and human action prediction, laying the groundwork for subsequent research in action recognition. Building upon Srivastava et al.'s work, Su et al. (2019) explored the fusion of autoencoders and the K-Nearest Neighbors (K-NN) model for predicting and clustering similar actions [1]. This hybrid approach capitalizes on the strengths of both autoencoders for dimensionality reduction and K-NN for classification, yielding enhanced performance in action recognition tasks. To further bolster the Encoder's resilience, the authors implemented Fixed Weights (FW) and Fixed States (FS), compelling the Decoder to rely on the Encoder's hidden states, thereby improving feature robustness. Another significant advancement was the integration of the Variational Autoencoder (VAE) with the Hidden Markov Model (HMM) [2][3]. This VAME (Variational Embeddings of Animal Motion) system learns probabilistic representations of 3D body key point sequences using a VAE, which are then modeled by an HMM to capture the temporal dynamics of actions. This approach facilitates the study of complex behaviors without needing labeled data, offering a valuable solution in scenarios where labeled data is scarce or costly to obtain.

Qin et al. (2022) developed a system for virtual hand control utilizing a Channel wise CNN (CW-CNN) regression model, incorporating Kalman filters to reduce latency, suitable for adapting to unsupervised tasks in industrial motion studies[9]. Leone et al. (2022) introduced a parallel classification method using logistic regression to simultaneously control multiple joint movements, enhancing the efficiency of categorizing complex motions in real time[10]. Further, Cheta et al. (2020) examined how training parameters influence the performance of neural networks in task recognition, which is vital for creating adaptable unsupervised classification systems[11]. Cho et al. (2020) combined different machine learning approaches, including a constrained autoencoder and support vector machines, to estimate finger force from surface electromyography (sEMG) signals, showing potential for broad applications in motion analysis[12].

In practical applications, Huang et al. (2020) demonstrated near-infrared spectroscopy for real-time monitoring in industrial composting, indicating the feasibility of such technologies for environmental monitoring[13]. Ravi et al. (2022) developed a semi-supervised quality control system using physics-based models and support vector machine (SVM) classifiers, suggesting enhancements in system robustness[14]. Yang (2024) utilized unsupervised learning techniques, such as principle component analysis (PCA) and K-means clustering, for analyzing deformation in built environments, offering methods adaptable for industrial motion pattern analysis[15].

Lastly, Shyamal and Swartz (2018) created a dynamic optimization-based system for electric arc furnace operations, integrating first-principles models to aid real-time decision-making[16]. These environments often present challenges such as variations in lighting, occlusions, and diverse human actions, which can hinder the performance of current algorithms. Future research should prioritize the development of algorithms that can address these challenges and enable accurate and reliable action recognition in real-time industrial settings.

III. Methodology

The proposed model uses an autoencoder architecture to extract meaningful representations from input sequence data then the K-means unsupervised learning is apply to group similar action. This architecture consists of an Encoder and a Decoder. The Encoder transforms the sequence of key points into a latent space representation using a Recurrent Neural Network (RNN) layer. The Decoder then uses this latent space input to reassemble the original input sequence.

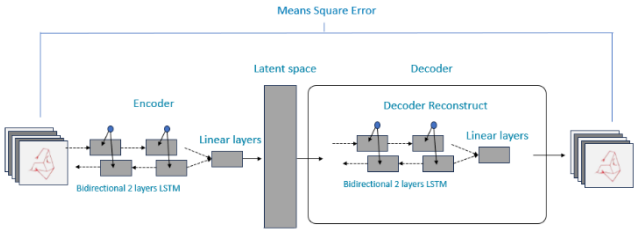

Fig 1: model architecture

Figure 1 illustrates the architecture of the proposed model, which uses the Encoder-Decoder to compress and rebuild the sequence action.

A. *Data collection*

This model uses the 2D human key points dataset to train the Encoder-Decoder model. The data was first extracted from a recorded video of shoe factory employees' activities during working hours. The worker in the video inspects the shoes' quality and stores them. The Mediapipe framework draws human body points and extracts the key points from the video, as shown in the figure below. When the worker is not in the camera area, we assign all the values of the human body key points to 0.

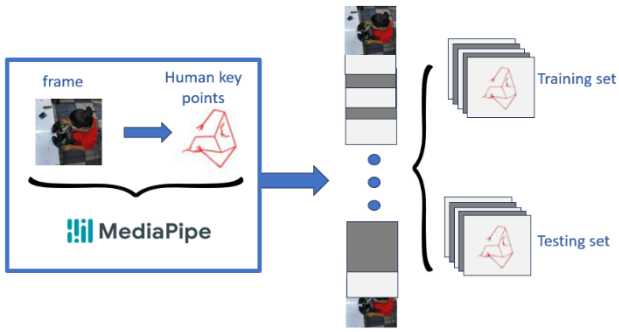

Fig 2:Data collection

B. *Data preprocessing*

The extracted keypoints from the video recordings are organized into a dataset with dimension of *(sequence length, feature size, x, y)* where *(sequence length)* represents the total number of sequence in the dataset, and $features\ size$ denotes the total number of keypoints. According to the MediaPipe framework, the feature size includes 33 keypoints.The (x, y) coordinates specify the positions of these keypoints in each frame. The (x, y) coordinates are subsequently flattened into a new dimension, resulting in a dataset with dimensions of *(sequence length, xy)*The min-max normalization technique is used to range the value of the dataset set from 0 to 1. Consider $K = \{k_{ij}\}$ as the raw keypoints dataset after flattening where:

- *i* ranges from 1 to total frame
- *j* range from 1 to 66 the number of features, which are the flatten *x* and *y* coordinates.

Let K_{min} be the minimum values of each keypoints joints and X_{max} is the maximum values of each keypoints joints. $K_{norm} = \{k_{ij}^{norm}\}$ represent the dataset after normalization, where each normalized element k_{ij}^{norm} is computed as described below:

$$K_{norm} = \frac{K - K_{min}}{K_{max} - K_{min}} \quad (1)$$

$$K_{min,j} = \min_{1<i\leq total\ frame} (k_{ij}) \text{ for j} = 1,2,\dots,66 \quad (2)$$

$$K_{max,j} = \max_{1<i\leq total\ frame} (k_{ij}) \text{ for j} = 1,2,\dots,66 \quad (3)$$

C. Autoencoder

The Encoder of our model takes sequences of input and tries to generate meaningful information at the last layer so that the Decoder can reconstruct the sequence input back from the latent space. Consider $E(x)$ is the Encoder with the sequence input of *x*, and z is the latent space in the bottleneck. Then, the relation between the Encoder and latent space is given by:

$$z = E(x) \quad (4)$$

The Decoder tries to reconstruct the input by taking the latent and rebuilding the sequence based on LSTM layers. Consider $D(z)$ is the Decoder and x' is the reconstructed sequence.

$$x' = D(z) \quad (5)$$

$$x' = D(E(x)) \quad (6)$$

The objective function is to minimize the error between the actual sequence and the reconstructed sequence. Denote f as the objective function that tries to minimize the error.

$$f(x) = \min (x - x') \quad (7)$$

D. Action clustering

The latent space from the Encoder exists in a higher dimension, which is not very suitable for visualization. The t-SNE algorithm projects the latent space from a higher-dimension space to a two-dimensional space, allowing for easier observation of the latent space's behavior, which shows similar properties. Once projected into a lower-dimensional space, unsupervised learning algorithm techniques become instrumental in grouping similar latent space representations. This paper proposes several unsupervised algorithms, including K-means, Hierarchical Clustering, Mean Shift, and DBSCAN algorithms. The optimal number of clusters for K-means was identified as 3 using the elbow technique, facilitating efficient data partitioning and revealing underlying patterns within the latent space. Additionally, through visual inspection of the results, as illustrated in the accompanying image, the efficacy of the clustering methodology in effectively capturing the data's inherent structure was confirmed.

IV. Experimental

The dataset being used is the recorded video. The experiment is set up in this environment by performing 3 actions: picking up, inspecting, and storing the shoes. The total duration of the video is 39 minutes and 47 seconds. The dataset is split into 70% training, 10% validation, and 20% testing. During training the Autoencoder, the hidden size value is set to 32, and the learning rate to 3×10^3, and the sequence length to 40. The model's Encoder comprises 2 LSTM layers to learn valuable information regarding key input sequence points.

Additionally, the bidirectional property of the LSTM layers enables the architecture to learn both the forward and backward directions of the input. During training, the size of the hidden state is 32, but due to the property set, the final output of the hidden state is 128. To ensure the hidden output matches the output size, the hidden layers are applied to the last hidden state of both directions of the LSTM layers, which is then fed into the Decoder. A setup has been developed where the Decoder mirrors the Encoder, creating a seamless process for data reconstruction. Using two layers of bidirectional GRU units in the Decoder enhances the ability to capture time-related patterns and context effectively. Afterward, a simple linear function is applied to align the key points of the reconstructed sequence with those of the input. This approach ensures that the reconstructed data faithfully represents the original, reflecting dedication to accuracy and meticulous research practices.

After completing our training phase, we will retrieve the latent space from the Encoder again. This latent space encompasses crucial information about the sequence of inputs. The unsupervised learning algorithm can be deployed using the extracted latent space to categorize similar key points into clusters. This method systematically groups data points based on their inherent similarities within the latent space.

V. Result

After training, the training loss between the input and the reconstruction key point is 0.0003, indicating that the model has effectively learned to reconstruct the input data with minimal error. To further analyze the model's performance, we apply the K-means unsupervised learning algorithm to the latent space, which helps cluster similar data points and uncover patterns in the dataset. This approach is compared with other unsupervised methods to ensure the robustness and accuracy of our findings. The elbow method is used to find the best value of k for the K-means algorithm. With the value of k of K-means equal to 3, the latent space is grouped as shown in the table below.

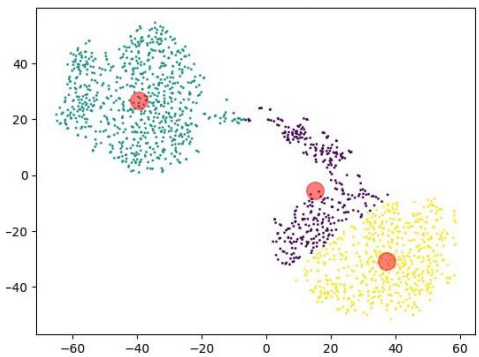

Fig 3: K-means clustering in the latent space

The model is then evaluated on an unknown dataset to assess its performance in settings resembling real-world conditions. The unseen dataset includes a range of actions comprising 8 minutes and 32 seconds of film. It consists of 36 separate operations, evenly split over three categories: 12 instances of picking up shoes, 12 instances of examining the shoes, and 12 instances of storing them in their respective locations. To effectively measure the model's performance, we use a confusion matrix, a useful tool for visualizing prediction accuracy across multiple categories. Using the confusion matrix, we calculate the classification accuracy, precision, recall, and F1 score for each type of action. This comprehensive evaluation explains the model's ability to generalize and accurately identify actions during training. By analyzing these metrics, we gain insights into specific areas where the model excels or may need further refinement, ensuring it performs robustly in practical applications. This process tests the model's effectiveness and highlights potential improvements that could enhance its applicability in dynamic, real-world environments.

Table 1: Confusion matrix

		Predict Class	
		Positive	**Negative**
Actual Class	**Positive**	TP=28	FP = 8
	Negative	FN = 8	TN=64

$$recall = \frac{TP}{TP + FP} = 77.77\% \tag{8}$$

$$Precision = \frac{TP}{TP + FN} = 86.66\% \tag{9}$$

$$Accuracy = \frac{TP + TN}{TP + TP + FN + FP} = 77.77\% \tag{10}$$

The model demonstrated robust performance. It recorded a recall rate of 86.11%, signifying that it accurately identified approximately 86.11% of the true action instances. The precision rate was 89.13%, indicating that around 89.13% of the model's positive action identifications were correct. Overall, the model achieved an accuracy of 86.11%. 86.11%These metrics collectively suggest that the model is highly effective at recognizing and correctly classifying specified actions with a low rate of false positives, which is critical in reducing operational errors within an industrial environment.

Further analysis of the model's performance across individual actions could provide insights into specific areas that may benefit from targeted improvements. The proposed model is compared with other traditional unsupervised machine learning models. This comparison aims to evaluate the effectiveness of the proposed model's feature processing and classification strategies relative to the conventional approach.

Table 2: Comparison of other unsupervised learning models

Metric	**K-means**	**DBSCAN**	**Mean Shift**	**Spectral Clustering**	**Hierarchical Clustering**
Accuracy	77.77%	44.44%	36.11%	61.11%	61.11%
Recall	77.77%	44.44%	36.11%	61.11%	45.83%
Precision	86.66%	45.33%	32.43%	68.31%	59.53%
F1 score	78.13%	40.77%	32.97%	57.57%	47.15%

Based on this table, the proposed method can have better accuracy in detecting human action without any label sets.

VI. Conclusion

This study presents an unsupervised learning methodology for human action recognition (HAR) in industrial contexts by using the Mediapipe framework and a sequence-to-sequence autoencoder model. This approach significantly enhances the scalability and practicality of time-motion studies by eliminating the dependency on labeled datasets, a common bottleneck in traditional HAR methods. Empirical evaluations substantiate the model's capacity to accurately classify worker activities in real time, enabling proactive operational management strategies to enhance efficiency and safety. Integrating t-SNE and K-mean clustering further reinforces the model's robustness.

This research shows promising outcomes in a controlled setting. However, further studies are needed to determine if these results can be applied to complex industrial environments with varying activities and conditions. The model's reliance on 2D skeletal key points could limit its effectiveness when objects block the view or the perspective changes significantly. The transition between different movements presents a significant challenge in this research. For instance, when a worker switches their movement from picking up an object to inspecting it, the model struggles to predict actions during these transitions accurately. Future research should include additional data types, such as depth information, to improve the model's robustness and accuracy in real-world conditions.

Despite these challenges, this study lays the groundwork for creating intelligent systems that can adapt and learn autonomously within their surroundings. The methodology introduced here holds the potential for real-time monitoring and analysis of human activities, which could significantly change how operations are managed in industrial contexts.

Acknowledgment

The Center of Excellence in Supply Chain Systems Engineering and Technology (COE LogEn Tech) at Sirindhorn International Institute of Technology, Thammasat University, fully supports this project. The first author acknowledges the scholarship awarded by Sirindhorn International Institute of Technology, Thammasat University, under the Thailand Advanced Institute of Science and Technology and Tokyo Institute of Technology (TAIST-Tokyo Tech) program. The scholarship is funded by the National Science and Technology Development Agency

(NSTDA) and the National Research Council of Thailand (NRCT).

SmartSense Industry is a leading innovator in industrial automation and smart technology solutions, leveraging IoT, AI, and data analytics to enhance industrial efficiency and sustainability. Their advanced sensor systems and predictive maintenance tools are revolutionizing various industrial sectors. The author would like to extend heartfelt gratitude to SmartSense Industry for sponsoring this project.

Chawalit Jeenanunta holds a B.Sc. in Mathematics and Computer Science and an M.Sc. in Management Science from the University of Maryland. He received his Ph.D. in Industrial and Systems Engineering from Virginia Polytechnic Institute and State University. He joined Sirindhorn International Institute of Technology, Thammasat University, Thailand, and is now an associate professor. He was a chair of the Management Technology curriculum, Head of the School of Management Technology, and Deputy Director for Building, Ground, and Properties. He is currently a Deputy Director for Academic Affairs and the head of the Center of Excellence in Logistics and Supply Chain Systems Engineering (LogEn).

REFERENCES

[1] K. Su, X. Liu, and E. Shlizerman, "PREDICT & CLUSTER: Unsupervised Skeleton Based Action Recognition," Nov. 2019.

[2] X. Zhang, D. Yi, S. Behdad, and S. Saxena, "Unsupervised Human Activity Recognition Learning for Disassembly Tasks," *IEEE Trans Industr Inform*, vol. 20, no. 1, pp. 785–794, Jan. 2024, doi: 10.1109/TII.2023.3264284.

[3] K. Luxem *et al.*, "Identifying behavioral structure from deep variational embeddings of animal motion," *Commun Biol*, vol. 5, no. 1, p. 1267, Nov. 2022, doi: 10.1038/s42003-022-04080-7.

[4] Y. K. Han and Y. B. Choi, "Human Action Recognition based on LSTM Model using Smartphone Sensor," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, Jul. 2019, pp. 748–750. doi: 10.1109/ICUFN.2019.8806065.

[5] V. de Oliveira Silva, F. de Barros Vidal, and A. R. Soares Romariz, "Human Action Recognition Based on a Two-stream Convolutional Network Classifier," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, Dec. 2017, pp. 774–778. doi: 10.1109/ICMLA.2017.00-64.

[6] N. Srivastava, E. Mansimov, and R. Salakhutdinov, "Unsupervised Learning of Video Representations using LSTMs," Feb. 2015.

[7] Z. Wang, R. Qin, J. Yan, and C. Guo, "Vision Sensor Based Action Recognition for Improving Efficiency and Quality Under the Environment of Industry 4.0," *Procedia CIRP*, vol. 80, pp. 711–716, 2019, doi: 10.1016/j.procir.2019.01.106.

[8] W. Vexo *et al.*, "Real-time Image Processing for Production Tracking in Manufacturing Plant," in *2023 Third International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*, IEEE, Jan. 2023, pp. 77–82. doi: 10.1109/ICA-SYMP56348.2023.10044734.

[9] Z. Qin, Z. He, Y. Li, S. Saetia, and Y. Koike, "A CW-CNN regression model-based real-time system for virtual hand control," *Front Neurorobot*, vol. 16, Dec. 2022, doi: 10.3389/fnbot.2022.1072365.

[10] F. Leone, C. Gentile, F. Cordella, E. Gruppioni, E. Guglielmelli, and L. Zollo, "A parallel classification strategy to simultaneous control elbow, wrist, and hand movements," *J Neuroeng Rehabil*, vol. 19, no. 1, p. 10, Dec. 2022, doi: 10.1186/s12984-022-00982-z.

[11] M. CHEŢA, M. V. MARCU, and S. A. BORZ, "Effect of Training Parameters on the Ability of Artificial Neural Networks to Learn: A Simulation on Accelerometer Data for Task Recognition in Motor-Manual Felling and Processing," *Series II - Forestry • Wood Industry • Agricultural Food Engineering*, vol. 13(62), no. 1, pp. 19–36, Jul. 2020, doi: 10.31926/but.fwiafe.2020.13.62.1.2.

[12] Y. Cho, P. Kim, and K.-S. Kim, "Estimating Simultaneous and Proportional Finger Force Intention Based on sEMG Using a Constrained Autoencoder," *IEEE Access*, vol. 8, pp. 138264–138276, 2020, doi: 10.1109/ACCESS.2020.3012741.

[13] Y. Huang, X. Sun, K. Liao, L. Han, and Z. Yang, "Real-time and field monitoring of the key parameters in industrial trough composting process using a handheld near infrared spectrometer," *J Near Infrared Spectrosc*, vol. 28, no. 5–6, pp. 334–343, Oct. 2020, doi: 10.1177/0967033520939323.

[14] D. Ravi, F. Barkhof, D. C. Alexander, L. Puglisi, G. J. M. Parker, and A. Eshaghi, "An efficient semi-supervised quality control system trained using physics-based MRI-artefact generators and adversarial training," *Med Image Anal*, vol. 91, p. 103033, Jan. 2024, doi: 10.1016/j.media.2023.103033.

[15] M. Yang, M. Li, C. Huang, R. Zhang, and R. Liu, "Exploring the InSAR Deformation Series Using Unsupervised Learning in a Built Environment," *Remote Sens (Basel)*, vol. 16, no. 8, p. 1375, Apr. 2024, doi: 10.3390/rs16081375.

[16] S. Shyamal and C. L. E. Swartz, "Real-Time Dynamic Optimization-Based Advisory System for Electric Arc Furnace Operation," *Ind Eng Chem Res*, vol. 57, no. 39, pp. 13177–13190, Oct. 2018, doi: 10.1021/acs.iecr.8b02542.