A Study on Efficient Multi-task Networks for Multiple Object Tracking

Xuan-Thuy Vo, Tien-Dat Tran, Duy-Linh Nguyen and Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering,
University of Ulsan
Ulsan (44610), South Korea
Email: {xthuy, tdat}@islab.ulsan.ac.kr; ndlinh301@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

***Abstract*—Multiple object tracking involves multi-task learning to handle object detection and data association tasks concurrently. Conventionally, object detection consists object classification and object localization (e.g., object regression) tasks, and data association is treated as a classification task. However, various tasks can cause inconsistent learning due to that the learning targets of object detection and data association tasks are different. Object detection focuses on positional information of objects while data association requires strong semantic information to identify same object target. Besides, advantageous character of multi-task learning is the correlation between tasks, and adopting such character in learning the networks can result in better generalization performance. However, existing multiple object tracking methods learn this information by treating multi-task branches independently. To understand the behaviours of multi-task networks in multiple object tracking, in this paper, we explore task-dependent representations through empirical experiments and observe that multi-task branches in multiple object tracking are complementary. To better learn such information, we introduce a novel Correlation Estimation (CE) module to estimate the correlation between object classification and bounding box regression based on statistical features of box regression quality. Finally, extensive experiments are conducted on the benchmark dataset MOT17. As a result, our method outperforms state-of-the-art online trackers without requiring additional training datasets.**

***Index Terms*—Multiple object tracking, multi-task network, multi-task learning**

I. Introduction

Multi-task learning is a learning paradigm [1], which learns the related information across multiple tasks to boost the generalization learning of all possible tasks. In the deep learning generation, multi-task learning encodes the task relatedness in two aspects: (i) network architectures with shared representation train multiple tasks simultaneously, (ii) task weighting is to balance the joint learning of multiple tasks to prevent an objective imbalance that one or more tasks can overwhelm training. Being multi-task learning problem, multiple object tracking (MOT) can be potentially improved from multi-task learning methods. Inspired by such ability, this paper takes two aspects of multi-task learning into account.

MOT is a basic yet challenging task in the computer vision research, and has been widely used in many applications such as object detection [2], video surveillance systems [3], human behaviors, and facial landmark detection. The MOT requires multi-task learning that learns the shared representation about: (i) object detection detects the presence of objects in all frames, (ii) data association associates these detections over the time-domain based on object identities. By definition, object detection task includes classification and regression subtasks, and data association is solved by the classification task. Accordingly, multi-task learning in MOT comprehends one regression task and two classification tasks. If these tasks are related, combining all tasks into a single tracking model is to learn the complementary information across tasks by using a shared layer mechanism. This strategy reduces the computation cost and boosts the generalization performance. Otherwise, if these tasks are unrelated, learning all tasks together without prior knowledge can degrade the performance. However, in the existing MOT methods [4]–[17], when jointly learning multiple tasks, they treat all tasks equally without investigating which tasks are related. Specifically, in two aspects of multi-task learning, these methods treat multi-task branches independently and balance task losses equally.

In the multi-task network aspect, most MOT methods [4]–[15], [17] design independent network branches for detection and data association tasks, which weakly learn the common features among tasks. These network designs increase model complexity and do not fully leverage the benefits of multi-task learning to the MOT task. To know how network branches work, in this paper, a comprehensive comparison between three tasks (two classifications and one regression) is conducted to investigate which tasks in MOT are related. By empirical experiments, we find the interesting fact that the three tasks in MOT are complementary, and regression in detection and classification in the data association can share the common information during training. This means bounding box regression is strongly correlated to the appearance features supervised by object identities for data association. Based on these insights, we propose a novel Correlation Estimation (CE) module to better learn complementary information between object classification and box regression according to quality features of box predictions.

II. Literature Review

Multi-task learning. In recent years, many methods have been proposed to improve the generalization performance of multi-task learning in deep network architecture [19]–[24]. Cross-Stitch [19] learns related tasks by linearly combining

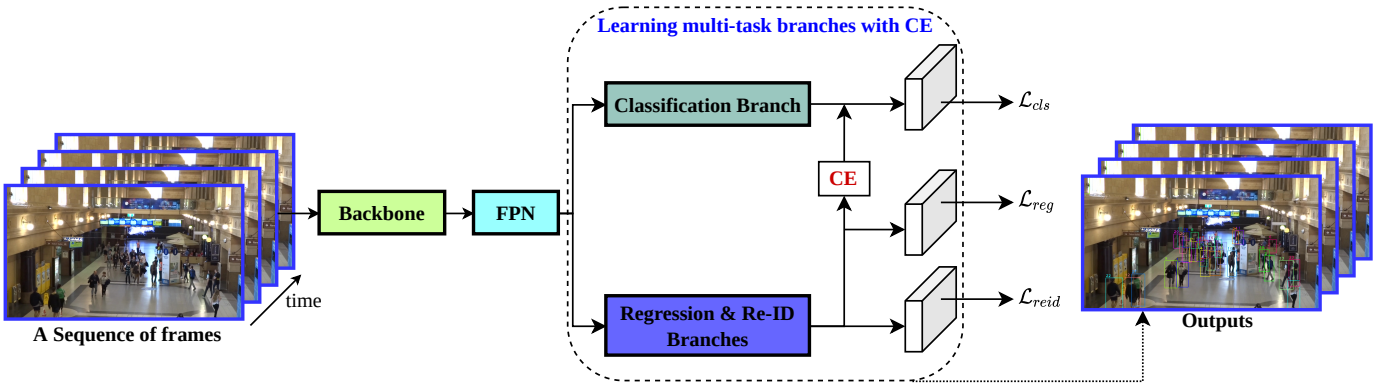

Fig. 1. The overall architecture of the joint-detection-and-tracking method consists of three parts: backbone network, feature pyramid network (FPN), and tracking head (multi-task branches). The video of this system is separated into a sequence of frames (at 30 frames per second) as the input images. The backbone network extracts informative features from the images. The feature pyramid FPN [18] indicates multi-level feature maps with different scales. The proposed multi-task branches with CE module (Correlation Estimation) include three branches in which classification and regression branches are sub-tasks of object detection, and Re-ID branch is used to predict identification scores for data association procedure. The output of the system is the coordinates of objects and identity numbers.

multi-task activation functions. MTAN [20] proposes a soft attention mask attached to the task-specific branch to learn task interactions. Fully-adaptive feature sharing [21] explores dynamic multi-task networks from thin to wide fashion based on the task grouping method. PAD-Net [22] assumes that learning auxiliary tasks can help target tasks, and final predictions are produced by gathering these auxiliary tasks via the multi-modal distillation approach. PAP-Net [23] empirically analyzes the multi-task network in PAD-Net and proposes the affinity module to learn the task relationship. MTI-Net [24] explores the task relatedness at different scales through three new modules, multi-scale multi-model distillation, feature propagation, and feature aggregation. Differently, this paper investigates the multi-task networks in terms of statistical features of box regression quality while existing methods explore the benefits of learning related tasks in different views such as multi-task activation functions [19], and attention mechanisms [20]–[24].

MOT. MOT is grouped into the online and offline methods according to the input frames. Online tracking methods use past and current frames as input images, thus reducing high computational costs. Offline tracking takes whole frames as input for the network. Even though offline methods bring significant improvements by combining motion features and optical flow, they rely on high model complexity. The first online tracking technique in [9], [25] consists detection and Re-ID (data association), based on CNNs. As MOT dataset [26] has object localization provided by detectors, for example, DPM, and Faster R-CNN [27], most of the tracking methods focus on data association procedure. Currently, several online one-shot trackers [13], [14], [16] join detection and Re-ID into a single end-to-end architecture to obtain a more efficient tracker, leveraging re-localization to enhance the data association step. This work uses the single end-to-end network as the baseline. Generic object detections [27]–[29] are applied for specific categories, such as human detections [26], which achieves remarkable improvements. This paper utilizes RetinaNet [28] for the detection step.

III. Learning Multi-task Branches in MOT

In this section, we discuss the task relatedness of MOT and propose efficient structures for learning multi-task branches. The proposed joint-detection-and-tracking network is described in Fig. 1. The used backbone network is ResNet-50 pre-trained on ImageNet for feature extraction. Following common methods, FPN [18] is used for constructing multi-level feature maps. We defer to the supplementary material the detailed dimensions of the backbone and FPN architectures. The tracking head (multi-task branches) with the proposed CE module learns the complementary information across tasks. In this paper, Re-ID appearance features are used for the data association task. The detailed architecture of the tracking head is shown in Fig. 2(f).

In the following, a thorough comparison between the detection and Re-ID network structures is performed to find the shared representation of the tracking head, shown in Table I. Each row in this table corresponds to each head structure in Fig. 2. Based on this comparison and its analysis, we propose the final multi-task branches with the CE module in Fig. 2(f).

TABLE I
Comparison of different types of head structure on the MOT17 validation set

Type	GFLOPs	#params (M)	MOTA↑	IDF1↑	MOTP↑
(a)	64.73	38.62	74.9	66.6	85.3
(b)	64.73	38.61	75.5	66.6	85.0
(c)	38.96	33.89	75.4	67.5	85.7
(d)	51.84	36.25	75.2	67.8	85.6
(e)	51.84	36.25	75.6	66.2	85.5
(f)	51.85	36.25	76.1	67.3	85.6

Three parallel branches have been widely explored in CenterTrack [13], JDE [14], and FairMOT [17] as shown in Fig. 2(a). This structure treats classification, regression, and Re-ID

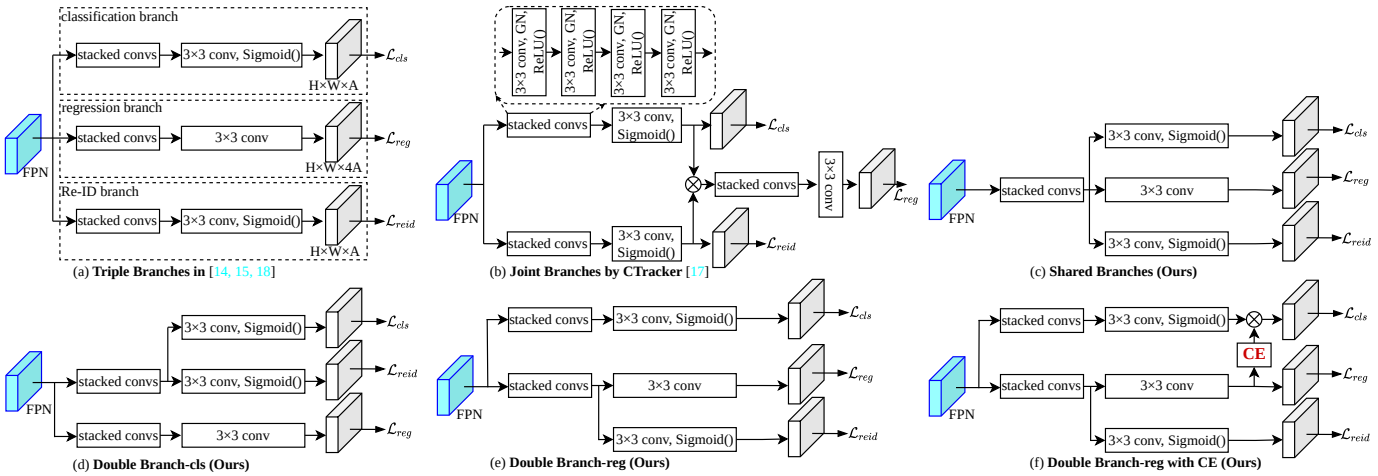

Fig. 2. Comparison between different types of tracking head: (a) Triple Branches used in CenterTrack, JDE, and FairMOT; (b) Joint Branches proposed by CTracker; (c) Shared Branches, where parameters of three branches are shared; (d) Double Branch-cls, where classification and Re-ID branches shares parameters; (e) Double Branch-reg, where regression and Re-ID branches shares parameters; and (f) Double Branch-reg with CE, which extends the double branch-reg structure by adding a correlation estimation (CE) module. $H \times W \times A$ denotes height, width, and the number of anchor boxes. $4A$ indicates four regressed offsets. $\bigotimes$ is element-wise matrix multiplication.

tasks independently, which has a high computational cost but inferior performance.

CTracker [16] utilizes the Joint Attention Module (JAM) to focus on local semantic features of the combined classification and Re-ID features, illustrated in Fig. 2(b). The regression branch uses combined features to improve detection and tracking performance. This method states that classification and Re-ID branches are complementary. Although this head structure takes advantage of task-dependent learning, it utilizes more stacked convolution layers causing computational overhead. Specifically, JAM achieves a MOTA score of 75.5% at 64.73 GFLOPs (Giga Floating-point Operations Per Second).

Fig. 2(c) describes the simplified structure of the tracking head in which parameters of three branches are shared. Interestingly, the performance is similar to type (b), while reducing the model complexity to 38.96 GFLOPs (by half of type (b)). It reveals that the three tasks are complementary. Thus, leveraging the correlation learning of the three tasks can improve tracking performance.

To consider how each task affects the others, the shared convolution is shown in Fig. 2(d), (e). More specifically, the classification and Re-ID branches have the same parameters to investigate related tasks across these two tasks, shown in Fig. 2(d). Alternatively, shared regression and Re-ID branches are performed in Fig. 2(e) to consider the collaborative learning of these two tasks. As a result, the MOTA score of type (e) is higher than type (d). It is understood that the classification task learns semantic features to distinguish objects and background, while the Re-ID task learns appearance features to identify two objects. Therefore, learning classification features complement Re-ID features. Re-ID and regression tasks make predictions on the same semantic features, and thus both can share the complementary information during training.

From the above observations, the extension of the double

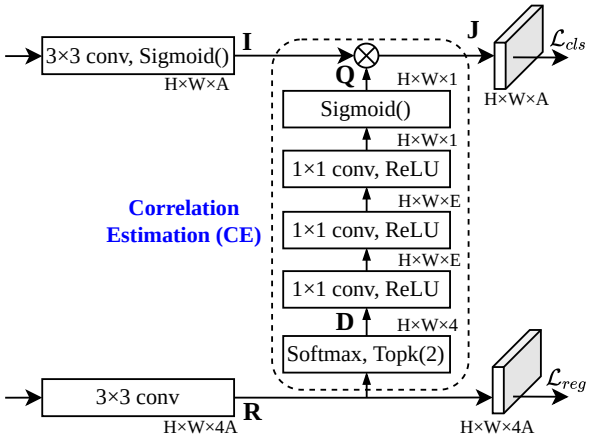

Fig. 3. The detailed sub-network of Correlation Estimation (CE), where E denotes the number of hidden channels. $\bigotimes$ is element-wise matrix multiplication.

branch-reg with the CE module is explored in Fig. 2(f) to learn three related tasks in an effective way. Motivated by the analysis in the PISA [30], the CE module is proposed to estimate the correlation between regression and classification tasks in a different perspective, regression quality, illustrated in Fig. 3. The regression distribution implemented by the softmax function φ is considered as Dirac delta distribution defined by the BBENet, which reflects ambiguities of the real dataset. The input of the CE sub-network is four offset parameters of the bounding box. Straightforwardly, the **Topk** values are used to measure regression quality **D**. If **Topk** values are higher, the bounding box distribution is sharper (i.e., corresponding to higher regression quality) and vice versus. These values guide classification score during NMS (Non-maximum Suppression). The CE module only has three 1×1

convs followed by ReLU and sigmoid to yield the quality feature $\mathbf{Q}$. Finally, the classification feature $\mathbf{I}$ is multiplied by the quality feature $\mathbf{Q}$ to leverage joint representation $\mathbf{J}$:

$$\mathbf{J} = \mathbf{I} \times \mathbf{Q}, \tag{1}$$

$$\mathbf{D} = Topk(\varphi(\mathbf{R})), \tag{2}$$

$$\mathbf{Q} = \delta(\sigma'_3(\mathbf{W'}_3\sigma'_2(\mathbf{W'}_2\sigma'_1(\mathbf{W'}_1\mathbf{D})))), \tag{3}$$

where $\mathbf{R}$ is the regression feature that denotes regressed offsets of the bounding box. $\mathbf{W'}_1 \in \mathbb{R}^{4\times E}$, $\mathbf{W'}_2 \in \mathbb{R}^{E\times E}$ and $\mathbf{W'}_3 \in \mathbb{R}^{E\times 1}$ are linear transforms implemented by 1×1 convolution.

The PyTorch code of the Correlation Estimation (CE) module is illustrated in the Algorithm 1. The input of the CE module is the regression features with four offset channels (box's center, height, and width). The selected **Topk** values must be suitable for input channel dimension, e.g., channel dimensions are divisible by **Topk** values. Thus, the **Topk** values can only be one, two, or four. During training and testing, we set **Topk** $= 2$ for all implementations since this value does not affect the performance.

Algorithm 1 Pytorch code of the CE sub-network

```
import torch
import torch.nn as nn
import torch.nn.functional as F

# E is the number of hidden channels
# topk_value forms regression quality
# (C_r is divisible by Topk value)

####initial_layers####
CE_net = nn.Sequential(
   nn.Conv2d(2*topk_value, E, kernel_size=1),
   nn.ReLU(inplace=True),
   nn.Conv2d(E, E, kernel_size=1),
   nn.ReLU(inplace=True),
   nn.Conv2d(E, 1, kernel_size=1),
   nn.Sigmoid())

def CE_module(regress_feat):
    # regress_feat (tensor): size [N, C, H, W]
    # N: batch size
    # C_r=4: number of regressed offset variables
    # H, W: height, width of feature map

    x = regress_feat
    N, C_r, H, W = x.size()

    # model distribution probability
    prob = F.softmax(x.reshape(N, 2, 2, H, W),dim=2)

    # quality estimation by Topk
    qe, _ = prob.topk(topk_value, dim=2)
    qe    = qe.reshape(N, -1, H, W)

    # forward to CE network
    corr_score = CE_net(qe)
    return corr_score
```

Model complexity is shown in the last row of Table I. The CE module only brings negligible additional GFLOPs, and thus it does not affect the training or testing time of the one-shot tracker. And the number of parameters (# params) is the same as type (e). Moreover, the extension of type (e) achieves a MOTA score of 76.1%, which surpasses all structures. It demonstrates the CE sub-network is simple yet effective.

IV. Experiments and Results

A. Datasets, Evaluation Metrics, and Implementation Details

The performance of the proposed method is evaluated on the benchmark dataset: MOT17 [26]. This dataset contain 7 training videos and 4 testing videos. More importantly, in this paper, we only train the model on the training set of the MOT17 while CenterTrack [13], JDE [14], and FairMOT [17] use combinations of other large-scale datasets for training. Thus, we do not include some methods in this paper for fair comparisons.

All results are measured by three standard metrics: Multiple Object Tracking Accuracy (MOTA), ID F1 score (IDF1) defined by CLEAR MOT, and Higher Order Tracking Accuracy (HOTA). Additional metrics include Multiple Object Tracking Precision (MOTP), the percentage of Mostly Tracked targets (MT), the percentage of Mostly Lost targets (ML), the total number of False Positive (FP), the total number of False Negatives (FN), and the number of Identity Switches (IDS). Among them, the MOTA score is the primary metric used for comparison with other methods.

All experiments are conducted by the deep learning Pytorch framework. The backbone ResNet-50 is pre-trained on the dataset ImageNet [31]. The weight initialization of the newly added convolutional layers in the FPN, tracking head, and CE module is filled from the normal distribution. Two GPU Tesla V100 devices with Cuda 10.2, and CuDNN 7.6.5 are used to train the model for 100 epochs with a batch size of 8. The Adam optimizer is applied for minimizing the detection and Re-ID objectives. The learning rate is set to $5\times e^{-5}$, and the number of anchor boxes tiled per one feature location is set to $A = 1$ for all implementations.

B. Results

This subsection analyzes the main performances conducted on the testing set of the benchmarks in subsection IV-B1, as well as the ablation study carried out on the MOT17 validation set in subsection IV-B2.

1) Comparison with State-of-the-art Methods: In this subsection, we describe the main results of our method on testing sets of the MOTChallenge benchmark, listed in Table II. The bold font indicates the best result among all state-of-the-art online methods. Since MOT benchmarks did not provide annotations for testing, all the detection and Re-ID results are uploaded and evaluated to the official MOT evaluation protocol.

Our proposed network achieves state-of-the-art performances on the dataset MOT17 in terms of the MOTA score and IDF1. More specifically, we achieve an MOT score of

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE TESTING SETS OF THE MOT17 BENCHMARKS

Dataset	Method	MOTA↑	IDF1↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓
MOT17	DMAN [4]	48.2	55.7	75.9	19.3	38.3	26218	263608	2194
	MOTDT [5]	50.9	52.7	76.6	17.5	35.7	24069	250768	2474
	Tracktor [12]	53.5	52.3	78.0	19.5	36.6	12201	248047	2072
	BLSTM-MTP [6]	53.6	55.8	-	23.5	34.4	23583	236185	**1845**
	Tracktor++ [12]	54.4	56.1	78.1	25.7	29.8	44109	210774	2574
	TADAM [8]	59.7	58.7	-	-	-	**9676**	**21629**	1930
	DeepSORT [9]	60.3	61.2	**79.1**	31.5	**20.3**	36111	185301	2442
	CenterTrack [13]	61.5	59.6	-	26.4	31.9	14076	200672	2583
	ArTIST [10]	62.3	59.7	-	29.1	34.0	19611	191207	2062
	SiamMOT [15]	65.9	**63.3**	-	**34.6**	23.9	18098	170955	3040
	CTracker [16]	66.6	57.4	78.2	32.2	24.2	22284	160491	5529
	Ours	**67.6**	57.6	79.0	31.7	26.0	16485	161502	4983

67.6%, which is superior to all the other trackers, including TADAM [8] (59.7%), DeepSORT [9] (60.3%), CenterTrack [13] (61.5%), SiamMOT [15] (65.9%), and CTracker [16] (66.6%).

TABLE III
INVESTIGATION OF E IN CE MODULE

E	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	#params
16	74.3	65.8	85.1	259	1954	0.34k
32	74.6	66.6	85.2	270	2060	1.18k
64	75.7	66.3	85.7	270	1493	4.41k
128	**76.1**	67.3	85.6	278	1668	17.02k
256	74.8	66.2	85.2	266	2115	66.81k

2) Ablation Study:

a) Hyperparameters in CE: Table III shows that the results of the model are sensitive to the variation of E. Specifically, setting $E = 128$ gets the optimal MOTA score among various values. Moreover, our CE module is very lightweight, which only takes 0.0004% of #params of the whole tracking network.

b) Performance on the MOT testing set: The detailed performances of our method on testing sets of MOTChallenge benchmarks are listed in Table IV.

TABLE IV
THE PERFORMANCE ON EACH VIDEO OF MOT17 TESTING SET

Video	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	IDs↓
MOT17-01	47.5	39.7	76.8	6	165	65
MOT17-03	87.2	66.8	78.9	124	3565	505
MOT17-06	56.6	56.2	78.5	65	397	216
MOT17-07	50.6	41.3	77.8	12	480	230
MOT17-08	30.3	30.1	83.1	11	225	178
MOT17-12	42.8	51.8	80.9	14	186	70
MOT17-14	40.1	43.1	78.4	13	477	397
Overall	67.6	57.6	79.0	735	16485	4983

c) Qualitative Results: The qualitative results of the proposed method are described in Fig. 4. Human identification is addressed by the identity number. Each curve denotes the predicted trajectory over the time domain.

V. CONCLUSION

This paper leverages the benefits of multi-task learning into improving the MOT network. The comprehensive analysis of the tracking head structure is investigated through empirical and theoretical analysis. As a result, we find the interesting fact that three tasks in MOT are complementary, and jointly learning such property can result in better generalization performance. To form better representation, the lightweight Correlation Estimation (CE) sub-network is proposed, which improves classification features by learning the estimated regression quality. The proposed method is evaluated on the dataset MOT17, achieving state-of-the-art performance. We hope that our method can serve as the simple baseline for multi-task learning research. In the future, the proposed method will be applied to multiple high-level tasks such as abnormal action detection, human pose tracking, and human behavior detection in video surveillance systems. It is a new and different perspective in solving multi-task learning and specific MOT task.

REFERENCES

[1] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[2] X.-T. Vo, L. Wen, T.-D. Tran, and K.-H. Jo, "Bidirectional non-local networks for object detection," in *International Conference on Computational Collective Intelligence*. Springer, 2020, pp. 491–501.

[3] X.-T. Vo, T.-D. Tran, D.-L. Nguyen, and K.-H. Jo, "Regression-aware classification feature for pedestrian detection and tracking in video surveillance systems," in *International Conference on Intelligent Computing*. Springer, 2021, pp. 816–828.

[4] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 366–382.

[5] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[6] C. Kim, L. Fuxin, M. Alotaibi, and J. M. Rehg, "Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9553–9562.

[7] G. Brasó and L. Leal-Taixé, "Learning a neural solver for multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6247–6257.

[8] S. Guo, J. Wang, X. Wang, and D. Tao, "Online multiple object tracking with cross-task synergy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8136–8145.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

Fig. 4. Qualitative results of the proposed method with some industrial surveillance videos.

[10] F. Saleh, S. Aliakbarian, H. Rezatofighi, M. Salzmann, and S. Gould, "Probabilistic tracklet scoring and inpainting for multiple object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 329–14 339.

[11] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi: Multiple object tracking with high performance detection and appearance feature," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–42.

[12] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 941–951.

[13] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European Conference on Computer Vision*. Springer, 2020, pp. 474–490.

[14] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 107–122.

[15] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe, "Siammot: Siamese multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 372–12 382.

[16] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *European Conference on Computer Vision*. Springer, 2020, pp. 145–161.

[17] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[19] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3994–4003.

[20] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[21] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5334–5343.

[22] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.

[23] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4106–4115.

[24] S. Vandenhende, S. Georgoulis, and L. V. Gool, "Mti-net: Multi-scale task interaction networks for multi-task learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 527–543.

[25] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[26] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[29] Z. Tian, C. Shen, H. Chen, and T. He, "Fully convolutional one-stage object detection2019," in *Conference: 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[30] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 583–11 591.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.