Abstract

Person re-identification (re-ID) is a challenging task in real-world. Besides the typical application in surveillance system, re-ID also has significant values to improve the recall rate of people identification in content video (TV or Movies). However, the occlusion, shot angle variations and complicated background make it far away from application, especially in content video. In this paper we propose an entropy based semantic feature alignment model, which takes advantages of the detailed information of the human semantic feature. Considering the uncertainty of semantic segmentation, we introduce a semantic alignment with an entropy-based mask which can reduce the negative effects of mask segmentation errors. We construct a new re-ID dataset based on content videos with many cases of occlusion and body part missing, which will be released in future. Extensive studies on both existing datasets and the new dataset demonstrate the superior performance of the proposed model.

1. Introduction

Person re-identification (re-ID), which aims to match people across different cameras with different viewpoints, is a very challenging task in real-world. Ubiquitous occlusion, complex background, illumination variations jointly bring the problem even harder. Most of the public datasets, such as Market-1501 [23], DukeMTMC [19, 29], are based on surveillance video, which empowers the person re-ID methods to achieve great progress in recent years. However, there is hardly any research working on person re-ID in TV or movie videos (which will be called content video in the following). The content video is a different domain from surveillance video. It is an artifact and manipulated by humans (cameraman or director) while the surveillance video captures the natural world with artificial time flow and constant view point. The artificial manipulations in content video will bring more people occlusions, body part missing, angle and scale variations.

An example of content video can be seen in Fig[1]. In Fig[1(a)] A and B are two consecutive shots describing the same scene but with different shot angles and camera distances. Matching the people in shot A to shot B can be regarded as a re-ID problem for half body and whole body. Re-identifying the person in the content video, especially matching the people in the shot with ill-conditioned for recognition or tracking to the other shot, which is ready to identify, is much more helping to improve the recall rate. Fig[1(b)] shows an example, where the detected people in shot A has frontal face and can be recognized. If we could re-identify the same person in shot B, we could recognize the character even if the person is side or back to the audience.

In this paper we want to solve the re-ID problem in both surveillance and content videos, focusing on the challenges of occlusion and body part missing. Those cases are much more remarkable in content videos. This problem is generally defined as partial re-ID or occluded re-ID, where the person in the probe or gallery is occluded or partially captured. Several methods [7, 8, 14] have investigated this problem and designed many methods to align the two peoples by static or dynamic local features. However, due to the lack of dedicated high-quality dataset for occlusion and body part missing, few methods have studied the alignment from semantic level in depth, not alone considering the uncertainty of semantic segmentations which would affect the alignment effects.

We propose a new re-ID model called entropy-based semantic alignment re-ID (ESA-ReID), dedicated for the oc-
Figure 1. Typical examples of content video. People in the content videos have large variations in angles, scales and occlusion.

Conclusion, missing body part issues in both content video and surveillance video. This model highly utilizes the detailed human body semantic feature and its uncertainty. We perform a semantic alignment of human semantic features based on visible score and confidence score defined on entropy. It will dynamically align the common and confidence semantic human regions without additional computation load. We also construct a new dataset called Drama-ReID, which, as far as we know, is the largest partial re-ID dataset retrieved from drama or movies. Extensive experiments on both public datasets and Drama-ReID reveal the advances of introducing both human body semantic features and its uncertainty. The main contributions of this paper are summarized as follows.

- A new person re-ID method for the partial and occluded re-ID problem is proposed, which investigates in depth the human semantic region information and achieves state-of-art in public dataset and Drama-ReID.

- A human segmentation task is associated to the re-ID model to design an entropy-based probability mask scheme, which split the feature map into confident and unconfident regions, providing more flexibility in semantic alignment.

- We introduce a semantic alignment using both semantic segmentation information and uncertainty information. The alignment automatically compares the common visible and confident semantic features.

- We construct a challenging large-scale re-ID dataset of content videos, which is by far the largest dataset focusing on partial and occluded person re-ID.

2. Related Works

Deep learning methods \cite{10,12,16,17,15,22} currently dominate the re-ID research community with significant performance on accuracy. Recent works \cite{22,23,16,3} further advance the state of the art on holistic person re-ID problem, through learning part-level features. For example, Sun et al.\cite{22} uniformly partitions the feature map to several local regions and learn a part-level representation respectively. Zhao et al.\cite{23} extracts part-level features by attention-based methods rather than grid cells or horizontal strips. Quan et al.\cite{16} utilizes a part-aware module to enhance the representation of body structural information. Fu et al.\cite{3} perform horizontal pyramid partition on feature map to extract enhanced discriminative information of all the scale-specific person parts. All those approaches\cite{22,23,16,3} have made some progress on holistic datasets\cite{24,19,29}. But an obvious problem with those methods is that their partitions of body are mainly spatial grid, rather than fine-grained and in terms of human semantic. Therefore, some methods\cite{20,15,10,17} based on semantic parsing are proposed. Song et al.\cite{20} and Qi et al.\cite{15} employ human mask information to facilitate person re-id models, where the mask-guided map can help to remove the background clutters. Kalayeh et al.\cite{10} utilizes an extra semantic segmentation network to harness local visual cues to learn the feature for each semantic piece. Then, they assemble the final discriminative representations with those semantic-level features. Quispe et al.\cite{17} combines the semantic parsing and saliency cues to improve the performance of person re-ID model. Although those methods employ the human semantic information, the representations of different semantic parts are not explicitly aligned and compared, which make them powerless to face the problems of missing parts and occlusion.

In order to solve the partial and occluded re-ID problems, some researchers have proposed several methods based on part matching\cite{27,7,8,21,14}. Zheng et al.\cite{27} proposes a local patch-level matching model called Ambiguity-sensitive Matching Classifier (AMC) and introduces a global part-based matching model called Sliding Window Matching (SWM). He et al.\cite{7,8} proposed a series of alignment-free methods that employ sparse feature reconstruction learning, called Deep Spa-
tial Feature Reconstruction (DSR)\[^7\] and Spatial Feature Reconstruction (SFR)\[^8\] respectively. Sun et al.\[^21\] proposed a visibility-aware part model (VPM) which learns to locate the visible regions through self-supervision on pedestrian images. VPM conducted a region-to-region comparison within their shared regions to suppresses the spatial misalignment. Miao et al.\[^14\] utilized pose landmarks as a guidance to construct the global feature and then combined the partial features for representation matching. Based on those related work, we could find that one of the main trends to solve the partial or occluded re-ID is to incorporate the additional information such as body pose or local parts. But one of the potential problems is that the local representations in those methods lack a correspondence with human fine-grained semantic parts, which is not aligned semantically. The other problem is the uncertainty of the addition information involved, such as the errors of the pose landmarks or the semantic segmentation. This problem will be more important when we incorporate the information in the multi-task way, which is more competitive in computation load. Our method will try to solve the issues by combining both the human semantic-level alignment and the semantic segmentation uncertainty.

3. Proposed Approach

Our proposed model consists of three components as shown in Fig.\[^3\] Similar to the traditional re-ID network, the input image will first be sent to the backbone network to obtain the basic features. Then a simple human parsing task is carried out based on the feature map of the backbone. This simple task will, on the one hand, regularize the backbone network to differentiate the different human body parts and the background, on the other hand, provide semantic segmentation for the following modules. The entropy-based masking module takes as input the semantic segmentation probability of human parsing task and the backbone feature map. Using the entropy of each semantic segmentation region, an entropy-based mask is generated, separating the feature map into confident features and unconfident features. We use the confident features in semantic part to do alignment while incorporating the unconfident features in the training part in an adversarial way to further boost the performance. Finally, the semantic alignment module aligns the semantic parts of the two people with both high visible probability and high confidence.

3.1. Backbone Network with Segmentation Task

We use ResNet50\[^6\] as our backbone network to extract basic feature maps of the given input image. We remove the average pooling layer and fully connected layer. Motivated by\[^22\], the stride of conv4_1 is set 1 to get larger feature map. Formally, we denote the feature map extracted from backbone as \( T \in R^{h \times w \times c} \), in which \( h, w \) and \( c \) denote the height, width, channel number respectively. We append a simple human parsing subtask following the feature map. This subtask takes as input the backbone output feature map \( T \) and then follows one \( 1 \times 1 \) convolution layer and a SoftMax layer. It outputs \( p(R_i|g) \), which denotes the probability of \( g \) belong to the human semantic body part \( R_i, i = 1, 2, ..., N \). \( N \) is the number of human semantic parts similar to the human parsing domain\[^4\]. Without loss of generality, \( N - 1 \) means the background. \( g \) denotes the pixel vector of \( T \) (there are \( h \times w \) number of \( g s \) in \( T \)), which is a \( c \)-dim vector. Parallel to the human semantic parsing task, we add another \( 1 \times 1 \) convolution after the feature map \( T \) to reduce the dimension in channels. This feature map, which is denoted by \( F \in R^{h \times w \times c_{new}} \), will be used as the following modules’ input.

3.2. Entropy-based Masking Module

As discussed in section\[^2\] we will utilize the human body semantic information to solve the occlusion or missing body part issues, which is also the main idea of several related works[10] [17]. However, none of them considered the effects of inaccuracy in human semantic segmentation. When the model depends on the result of segmentation task, the small inaccuracy at the feature map of higher layer might be enlarged in the raw image. Here we introduce an entropy-based mask to get both confident and unconfident feature part. Both of them will provide more flexibility for the following network to utilize the features to align the features semantically.

The human segmentation task would output the probability \( P(R_i|g) \) of each pixel in the feature map \( T \) belonging to semantic part \( R_i \). However, we could not guarantee the accuracy of this task, especially during the training process. To evaluate the uncertainty of the segmentation result, we use the entropy of segmentation results for \( g \), denoted by \( E_g \) as follows:

\[
E_g = - \sum_{i=0}^{N} p(R_i|g) \log p(R_i|g) \tag{1}
\]

If \( E_g \) is small, it means that the model can identify the corresponding part of human body with low uncertainty. In other words, \( g \) would be the true semantic features predicted by \( P(R_i|g) \) with high confidence. \( E_g \) is bounded by \( E_{max} = \frac{1}{N} \log N \). Then the entropy mask \( M_g \) for unconfident feature can be generated by choosing fixed threshold \( \tau \) as follows:

\[
M_g = \begin{cases} 
E_g/E_{max}, & E_g/E_{max} \geq \tau \\
0, & \text{otherwise}
\end{cases} \tag{2}
\]

Oppositely, we could also get the confident semantic attention map \( \hat{M}_g \) as

\[
\hat{M}_g = 1 - E_g/E_{max} \tag{3}
\]
Performing an element-wise multiplication on the feature map $F$ with both entropy mask $M$ and entropy attention map $\tilde{M}$, we could get the unconfident entropy features and high distinctive semantic features respectively, which is shown in Fig.3(a). This entropy-based mask plays an important role in distinguishing reliable semantic features as well as in the following alignment module. In general, the unconfident part generally involves the human body intersection part and human profile, since the human parsing model often confuses around the human’s nearby body parts or human boundary to background. Thus, by masking out the unconfident feature map, the human semantic features will be more accurate and helpful for the semantic alignment for people with different visible body part in re-ID. The entropy-based mask scheme also helps in the model training process, which will be discussed in next session. Performance comparison and visualization can be found in Section 4.4 and 4.5.

### 3.3. Semantic Alignment

This section describes the semantic alignment part, which will take each semantic part feature (exclude the background) to form the final feature for re-ID. Note that for the occlusion case in re-ID, the figure pair is not aligned,
i.e. the persons in the image pair have different body parts visible. Thus, the alignment should use the features corresponding to the shared body parts. As mentioned in Section 2, some work uses fixed partition of feature map, which is coarse. Other works, although utilize the semantic information, neglect the uncertainty of semantic segmentation which might affect the final performance.

Now we have semantic body part features based on the feature map $F$ and semantic segmentation probability $P(R_i|g)$. Then denote the local features of each semantic region as $f_i$, which is calculated by

$$f_i = \sum_{g \in R_i} p(R_i|g)gf, \forall i \in \{1, 2, ..., N\} \tag{4}$$

where $gf$ is the pixel features in $F$ spatially corresponding to the $g$ in feature map $T$. $F$ and $T$ have same spatial dimensions. It is the weighted sum of all the candidate pixels in $F$ using probability denoted by "WP" (weighted pooling) in Fig 2. Note that Eq 4 does not consider the segmentation uncertainty, i.e. entropy mentioned in Section 7. We further extend the Eq 4 as follows

$$\tilde{f}_i = \sum_{g \in R_i} \tilde{M}_g p(R_i|g)gf, \forall i \in \{1, 2, ..., N\} \tag{5}$$

which means we enhance the confident semantic features and suppress the unconfident semantic features. Besides, we introduce the visible score $S_i$ similar to Eq 4 for each region defined by $S_i = \sum_{g \in R_i} P(R_i|g)$. If a human semantic region is invisible, occluded or non-significant, the visible score will be very small. Combining both the confident local features and visible score, we could formulate the distance of the persons in image $p$ and $q$ with semantic alignment as

$$d_{p,q} = \sum_{i=1}^{N-1} S_i^p \cdot S_i^q \cdot D(\tilde{f}_i^p, \tilde{f}_i^q) \tag{6}$$

where $D$ is a distance metric; $N-1$ means we do not select the background region indexed by $N$. We can see that if any semantic part is unavailable in either of $p$ or $q$, $S_i^p \cdot S_i^q$ will be close to zero, thus the distance will not consider that part.

We also define the unconfident part feature as $f_{un}$, which can be obtained by multiplying $M_g$ on feature map $F$ with global average pooling denoted by "GAP" in Fig 2. We treat $f_{un}$ as the feature for a special semantic region, which contains the uncertain semantic region. This type of region, which will be shown in Section 3.2, is generally the intersection of body parts and profile. Analogue to the semantic region, define the confident score $S_{un}$ for this region as $S_{un} = \sum_g M_g$, we could get an extended generic distance $\tilde{d}_{p,q}$ combining both the confident features and unconfident feature:

$$\tilde{d}_{p,q} = \sum_{i=1}^{N-1} S_i^p \cdot S_i^q \cdot D(\tilde{f}_i^p, \tilde{f}_i^q) + S_{un}^p \cdot S_{un}^q \cdot D(f_{un}^p, f_{un}^q) \tag{7}$$

Adding the unconfident region feature into the distance metric will play an adversarial regularization role, which help to stabilize the training and improve the final performance. At the beginning of training, we could not get accurate human parsing result. The model is more dependent on the unconfident feature, which might include almost all features of the feature map. As the training going on, the human parsing task will be more accurate and the semantic part will be enforced by the supervision signal. The meaningful semantic region with high certainty will gradually dominate the distance metric. It’s like a self-adversarial game that maintains a dynamic balance between semantic features and high entropy features. We will investigate it experimentally in Section 4.4.

### 3.4. Network Losses

Several losses are defined considering the complicate network structure with both entropy-based mask and semantic alignment. In principle, we use traditional identification loss (ID loss) for each semantic region (including the unconfident region defined above) and triplet loss on the entropy based semantic aligned distance. For each semantic region $R_i$, except the background, we append the local feature $f_i$ for $R_i$ with one FC layer and one SoftMax layer to predict the people ID $\hat{y}_i$. The ID loss of semantic regions is defined as

$$L_{ID} = \sum_i^{N-1} S_i \cdot CE(y, \hat{y}_i) \tag{8}$$

where $CE$ is the cross-entropy loss for people id classification, $S_i$ is the visible score for semantic region $i$. Similarly, we also append one FC layer and one SoftMax layer after the unconfident features $f_{un}$. The extended loss considering the unconfident semantic region is

$$\tilde{L}_{ID} = \sum_i^{N-1} S_i \cdot CE(y, \hat{y}_i) + S_{un} \cdot CE(y, \hat{y}_{un}) \tag{9}$$

where $\hat{y}_{un}$ is the ID prediction based on the unconfident feature. Note that confident scores for each semantic region even including the unconfident semantic region are involved in those loss definitions. Also, for the confident regions, we use $f_i$ defined in Eq 4 rather than $\tilde{f}_i$ defined in Eq 5 expecting the loss has strong supervision on the semantic segmentation.

For the triplet loss[9], it is defined as

$$L_{triplet} = \max\{0, \tilde{d}_{a,p} - \tilde{d}_{a,n} + m\} \tag{10}$$
where $a$ and $p$ are the positive pair while $a$ and $n$ represent negative pair; $d$ is the extended distance defined in Eq. 7; $m > 0$ is the margin the triplet loss wants to guarantee between positive and negative pairs. The final loss is written as

$$L_{all} = \lambda L_{parsing} + \tilde{L}_{ID} + L_{triplet}$$

where $L_{parsing}$ is the general segmentation loss for the human segmentation task, which is not expanded in detail for brevity. $\lambda$ is the weight for the human segmentation task. Because human parsing task is important for our model, we will investigate the chosen of $\lambda$ in ablation study.

4. Experiments

4.1. Datasets and Evaluation Measures

To demonstrate the performance of our method on re-ID problem with occlusion, missing part in both traditional surveillance and video content case, we evaluate the proposed model on dataset with both partial re-ID dataset and holistic re-ID datasets. However, there is few large data for re-ID problem with occlusion, missing part, especially in video content domain. We construct our own data set called Drama-ReID.

Drama-ReID. Drama-ReID is a dataset in content video domains, i.e. the human figures are extracted from the TV and movies. As we discussed before, the content video is artifact, which is different from the traditional surveillance case. The people in the figure will be captured with different shot languages such as close-up shot, full-shot, different angles, etc. We extracted $\sim$300,000 images of $\sim$10,000 identities from $\sim$500 different TV shows with existing human Yolo detector [18]. Each identity has at least 30 pictures with different poses, shot angle, half-full bodies, etc. Fig. 4 shows several examples of Drama-ReID. Actors with different clothing will be treated as different identities. As far as we know, this dataset is now the largest re-ID dataset in both content video and surveillance video. We will release it in the near future to inspire research insights on complicated re-ID cases with occlusions, missing parts, angles, poses scale variations, etc.

Partial-REID [28] and Partial-iLIDS [26]. Partial-REID and Partial-iLIDS are typical evolution dataset for occlusion and missing part research in re-ID. Partial-REID includes 600 images from 60 people, with five full-body images and five partial images per person. We follow the evaluation protocols in [25] where 300 full-body images of 60 identities are used as the gallery set and 300 occluded-body images of the same 60 identities are used as the probe set. Partial-iLIDS is derived from iLIDS [26], which is collected in an airport and the lower-body of a pedestrian is frequently occluded by the luggage. The iLIDS dataset contains a total of 476 images of 119 people captured by multiple non-overlapping cameras.

Market-1501 [24] and DukeMTMC-reID [19, 29]. We also evaluate our model in public full-body re-ID dataset. Market-1501 contains 1501 identities, which is extracted from 6 camera viewpoints, 19732 gallery images and 12936 training images detected by DPM [2]. DukeMTMC-reID contains 1404 identities, 16522 training images, 2228 queries, and 17661 gallery images.

Evaluation Metrics. We use Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP) to evaluate the quality of different person re-ID models. All the experiments are performed in a single query setting. However, for Drama-ReID, we also require the evaluation metric PR-AUC [5], which is defined as the area under the P/R curve. Because the re-ID problem is in essence a binary classification problem, PR-AUC will give us more insights about the model performance than ranking metric. What’s more, in real application, we commonly use re-ID to matching tracking results in different shot in content video. Thus, AUC provides a guidance to select the threshold to tradeoff between precision and recall.

4.2. Implementation Details

Our backbone is ResNet50 [6] without the average pooling layer and fully connected layer. At the same time, the stride of conv4_1 is set to 1 in order to obtain larger feature map. We initialize the backbone model by the ImageNet [11] pre-trained model. In our experiment setting, the input image is resized to 384 $\times$ 128 and augmented by random flipping and random erasing [30]. We use batch hard triplet loss [2] and set batch size to 64 with 8 identities (8 images per identity). We use standard Stochastic Gradient Descent (SGD) optimizer, with the initial learning rate being 0.1 and decaying to 0.01 after 30 epochs. The proposed method is trained with 70 epochs. Before training, we process the images using human parsing model [4], which gives 20 semantic labels. By combing the minor human semantic part, we get 8 typical parts, namely background, head, torso, upper-arm, lower-arm, upper-leg, lower-leg and foot. At the same time, we also train a strong baseline model which only has the Resnet50 followed by the hard triplet loss.

4.3. Result Comparisons

4.3.1 Results on Drama-ReID.

Table 1 shows the result of our model compared with several existing works. Our model achieves 96.7% Rank-1 accuracy and 91.3% mAP, which outperforms all the previous methods. At the same time, to compare the PR-AUC performance with existing work, we also trained some models following the authors’ implementations and official open-source code on Drama-ReID. The PR-AUC of our model is 0.86, which is much better than them.
Table 1. Results on Drama-ReID dataset

| Method               | Rank1 | mAP  | PR-AUC |
|----------------------|-------|------|--------|
| DSR(CVPR2018)        | 82.5  | 73.0 | 0.47   |
| SFR                  | 86.9  | 76.3 | 0.60   |
| PCB(ECCV2018)        | 93.8  | 84.5 | 0.79   |
| VPM(CVPR2019)        | 94.5  | 86.7 | 0.81   |
| Baseline             | 93.2  |      |        |
| ESA-ReID(ours)       | 96.7  | 91.3 | 0.86   |

Table 2. Results on partial re-ID datasets

| Method     | Partial-REID | Partial-ILIDS |
|------------|--------------|---------------|
|            | Rank1 | rank3 | Rank1 | rank3 |
| MTRC       | 23.7  | 27.3  | 17.7  | 26.1  |
| AMC        | 37.3  | 46.0  | 21.0  | 32.8  |
| DSR        | 50.7  | 70.0  | 58.8  | 67.2  |
| SFR        | 56.9  | 78.5  | 63.9  | 74.8  |
| STNReID    | 66.7  | 80.3  | 54.6  | 71.3  |
| VPM        | 67.7  | 81.9  | 65.5  | 74.5  |
| PGFA       | 68.0  | 80.0  | 69.1  | 80.9  |
| Baseline   | 55.0  | 76.1  | 56.2  | 66.3  |
| ESA-ReID(ours) | 78.6  | 84.3  | 73.6  | 82.4  |

4.3.2 Results on Partial-REID [28] and Partial-ILIDS [26].

To compare the performance on partial re-ID case, we compare our model with several existing partial person re-ID methods, including MTRC [11], AMC+SWM [27], DSR [7], SFR [8], STNReID [13], VPM [21] and PGFA [14]. Same to the previous works, we train our model on Market-1501 training set. As shown in Table 2, the performance of our model is quite competitive. The Rank-1/Rank-3 of our method achieves 78.6%/84.3% and 73.6%/82.4% on Partial-REID and Partial-ILIDS. Comparing to the strongest competing method PGFA [14], our method surpasses it by 10.6% on Partial-REID and 4.5% on Partial-ILIDS in Rank-1 respectively, which are a large margin. The results demonstrate that the entropy-based semantic feature alignment is very useful for partial and occlusion re-ID.

4.3.3 Results on Market-1501 [24] and DukeMTMC-reID [19].

We also test our model on Market-1501 and DukeMTMC-reID. As shown in Table 3, even on holistic person re-ID datasets, our model achieves comparable performance with state-of-the-art.

4.4. Ablation Study

In this section, we do ablation study on several components in our proposed model.

4.4.1 The Impact of Coefficient $\lambda$.

The loss function defined in Eq 11 has a hyper parameter $\lambda$. It’s the weight on human semantic parsing task, which is important in our task. We conduct a series of experiments with different $\lambda$s to select the best choice. As shown in Fig 5 when $\lambda$ is small (for example 0.01), the performance of semantic parsing task is difficult to improve, because incorrect semantic matching will reduce the final performance of the model. When $\lambda$ is too large, the optimizer will favor too much on semantic parsing task rather than semantic alignment and matching task, which will also lead a poor performance on the representation learning. For the datasets Partial-REID and Partial-ILIDS, we best $\lambda$ is our
4.4.2 The Impact of Threshold $\tau$.

In Eq.2 and Eq.3, $\tau$ is the threshold of the entropy mask. It works like a gate. During each train step, it will select the confident parts in the semantic segmentation region. As shown in Fig.6, we increase $\tau$ from 0.1 to 0.9. When $\tau$ is small, such as 0.1, the gate will keep large number of unconfident features. It has high chance to lose useful human semantic features and the unconfident feature will be overwhelmed, which wont help for the final result. On the other side, when $\tau$ is large, many regions with high uncertainty will be involved in the semantic regions, degrading the alignment performance based on the semantic regions.

4.4.3 The Impact of Unconfident Features.

We further investigate the performances if we remove or change the unconfident features. First, we replace the unconfident feature by global feature, i.e. removing the entropy-based mask and the unconfident features in the compare distance in Eq.7. The model actually degenerates to a model that does alignment with semantic segmentation feature and the global feature. We denote the result as ”g” in Table4. Second, we remove the unconfident feature part and only use semantic segmentation features, whose result is denoted by ”w”. Third, we design a dynamic threshold experiment in entropy-based mask generation. We only choose features larger than the largest half part of entropy as the unconfident region, whose result is denoted by ”d”. By comparing the results of ”g” and ”d” with our model, we could find the advantage of the unconfident region features by the entropy mask. The results demonstrate the adversarial regularization effects of incorporating unconfident features into the alignment distance as mentioned in Section2. The incorporation of unconfident features in semantic alignment will stabilize the training process and boost the final performance in a self-adversarial way. The results of ”d” and ESA-ReID indicate that a fixed $\tau$ is preferable under this self-adversarial regularization settings.

4.5. Visualization

To see more clearly the entropy value of different regions, we visualize the semantic parsing result and entropy score of the person images. As shown in Fig6, we found that the high entropy region appears mainly at the edges of human semantic parts and intersection of semantic regions. Those regions are difficult to determine exactly which semantic part they belong to. If we force those regions simply by probability to certain semantic part regions, those incorrectly regions will contaminate the semantic features. On the other side, we can see that the unconfident part has some other semantic part which is not common, such as the 5th persons shoes and the 7th persons hat. Those minor but distinguishable region is not involved in the general semantic regions. But by incorporating the entropy-based mask and unconfident region feature, those features can also be utilized in our model.

5. Conclusions

In this paper we propose an entropy-based semantic alignment model. This model utilizes the human semantic region information by a simple human parsing task. The inaccuracy of human semantic parsing is innovatively utilized
Figure 7. Visualization of the semantic parsing result and entropy score of person images. The 1st row shows input person images; the 2nd row is the corresponding entropy score map; the 3rd row is the corresponding semantic parsing results.

by an entropy-based mask scheme. The semantic alignment will consider both the visible score and confident score of each semantic part, which is expected to handle the case of occlusion and partial re-ID. We also construct the largest re-ID dataset especially for the content videos, which has many cases of partial, occlusion, shot angle variations. Our model shows prominent results on both our new dataset and existing public datasets. However, there is still many other unsolved cases such as angle variations, large pose-variations in re-ID, which can be seen in our new dataset. We will release the dataset, hoping to inspire more insights on the re-ID in the domain of content video.
References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9):1627–1645, 2009.

[3] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 8295–8302, 2019.

[4] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 932–940, 2017.

[5] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology, 143(1):29–36, 1982.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[7] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7073–7082, 2018.

[8] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. arXiv preprint arXiv:1810.07399, 2018.

[9] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

[10] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1062–1071, 2018.

[11] Shengcai Liao, Anil K Jain, and Stan Z Li. Partial face recognition: Alignment-free approach. IEEE Transactions on pattern analysis and machine intelligence, 35(5):1193–1205, 2012.

[12] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4099–4108, 2018.

[13] Hao Luo, Wei Jiang, Xing Fan, and Chi Zhang. Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Transactions on Multimedia, 2020.

[14] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 542–551, 2019.

[15] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. Maskreid: A mask based deep ranking neural network for person re-identification. arXiv preprint arXiv:1804.03864, 2018.

[16] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 3750–3759, 2019.

[17] Rodolfo Quispe and Helio Pedrini. Improved person re-identification based on saliency and semantic parsing with deep neural network models. Image and Vision Computing, 92:103809, 2019.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[19] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision, pages 17–35. Springer, 2016.

[20] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1179–1189, 2018.

[21] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 393–402, 2019.
[22] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[23] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017.

[24] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[26] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR 2011*, pages 649–656. IEEE, 2011.

[27] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[28] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4678–4686, 2015.

[29] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[30] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.