# Domain Adaptation for Violence Detection in Camera Surveillance

Group 6 & Group 14

A

# Motivation

# Motivation: Domain Gap in CCTV

**The Background**

- CCTV (Camera surveillance) shown to be **effective at crime prevention**
- Automatic **Violence Detection** required for large-scale adoption

**The Challenge**

- Each camera installation presents new domain
- Few instances of crime captured by any given camera
- Laborious to label

We require approaches to improve performance without **labeled training data**

# Unsupervised Domain Adaptation: Improve classification performance on real-world CCTV dataset using only labeled Hockey dataset

**Source Dataset: Hockey**

- Dataset of hockey games
- Labels available
- 50% fight, 50% non fight

**Target Dataset: UCF**

- Dataset of real-world CCTV
- No labels available
- Dataset with bias towards non violence
- Extra difficulty: very diverse conditions (lighting, quality, …)

**The Goal:**
Achieving good classification accuracy on UCF

**The Problem:**
H**ockey-finetuned** model performs **badly** on UCF due to **large domain gap**

**Our Task – Unsupervised Domain Adaptation:**
Investigating approach to **improve UCF performance** using only **labeled source** data and **unlabeled target** data

B

# Methodology

# Methodology

**Method 1 – UDA:**

- Aligning latent representations between source and target domain

**Method 2 – SSL:**

- Improving feature representation for target domain through maximum entropy coding

**Key Contribution – Combination:**

- UDA and SSL can be combined to achieve significant performance gains on the target domain

**Overview Experiments**

Unsupervised Domain Adaptation

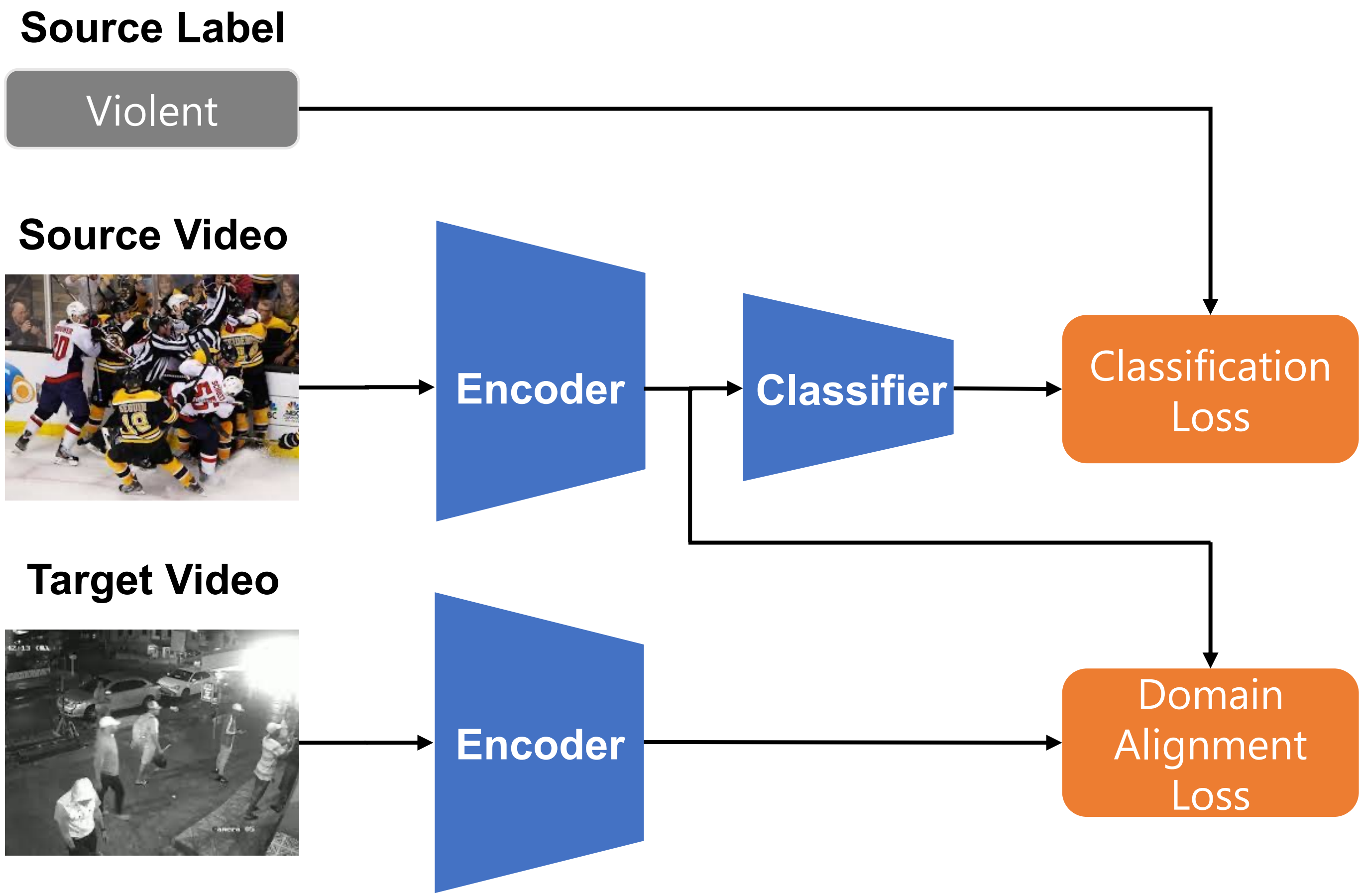
Self Supervised Learning

**UDA + SSL**
**=**
**Significantly Improved Performance**

C

# Experiments

# Unsupervised Learning To Align Latent Representation Between Source and Target Domain




**The Approach:**

- Good performance on source-domain due to labeled training data
- Align distribution of source and target domain in feature space

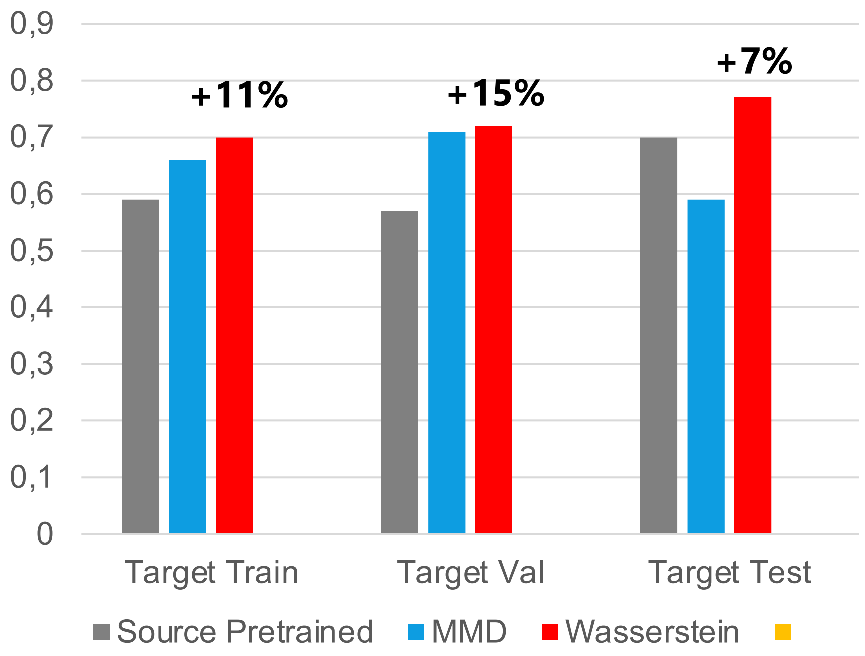
→ Good **performance on source translates** to good **performance on target**

**How to Align Domains:**

1. **Adversarial Approach**: Good performance but very **unstable**
2. **Discrepancy approach:** Investigated a) **Maximum Mean Discrepancy** and b) **Wasserstein** Distance

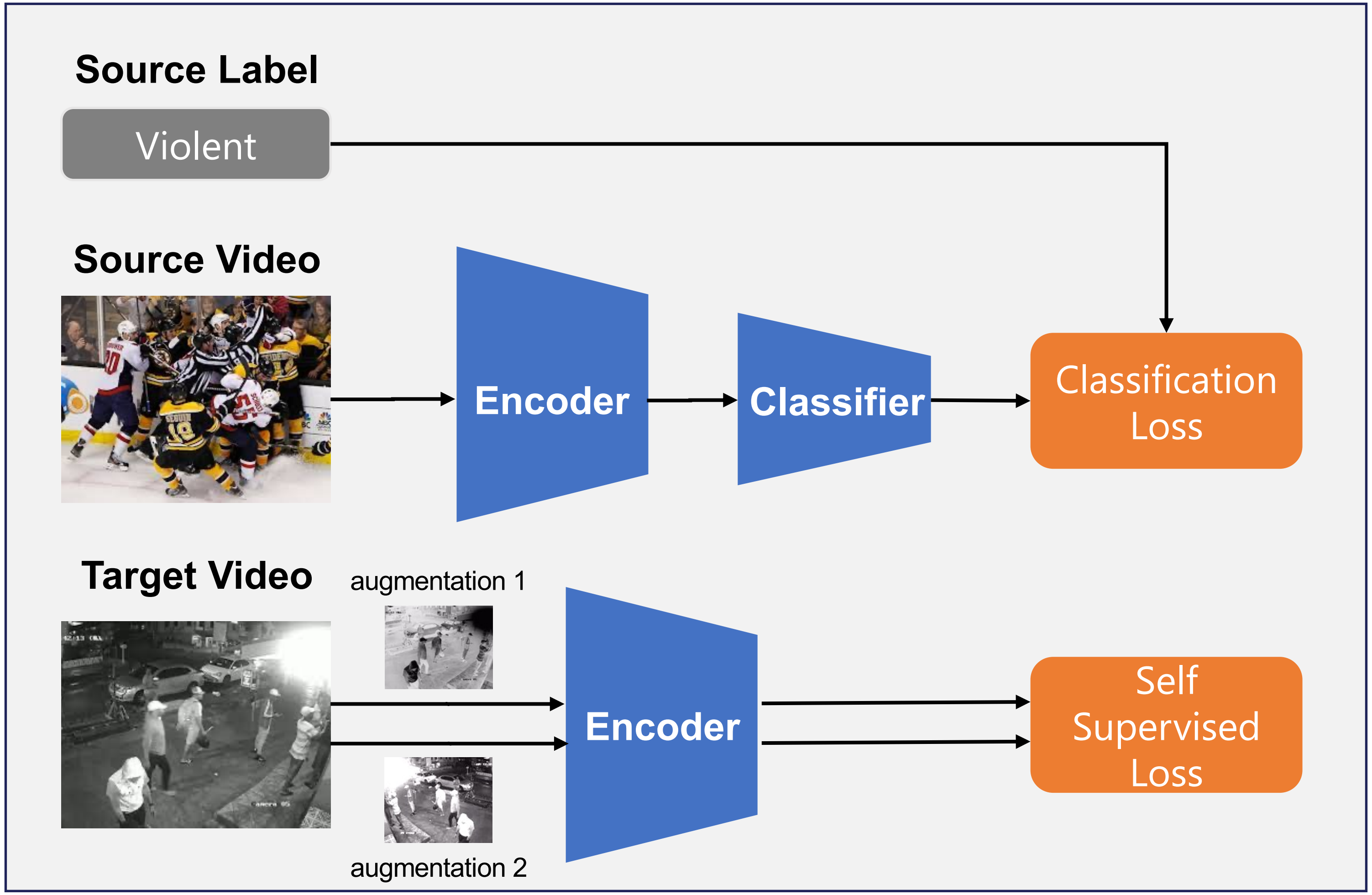
# Unsupervised Domain Adaptation Effective At Improving Target Domain Performance

**Performance Values Train, Validation, and Test Set**
[All values measured in % accuracy]




- UDA **effective** and **improves accuracy on all datasets**
- **Wasserstein** Loss more **effective** than **Maximum Mean Discrepancy**
- **Ablation:** Best performance when average pooling time dimension. Hypothesis:
  - Simplified latent space easier to align
  - Videos are uniformly violent → time information not that important

# Self-Supervised Learning To Learn Good Representations for Target Domain




**The Approach:**

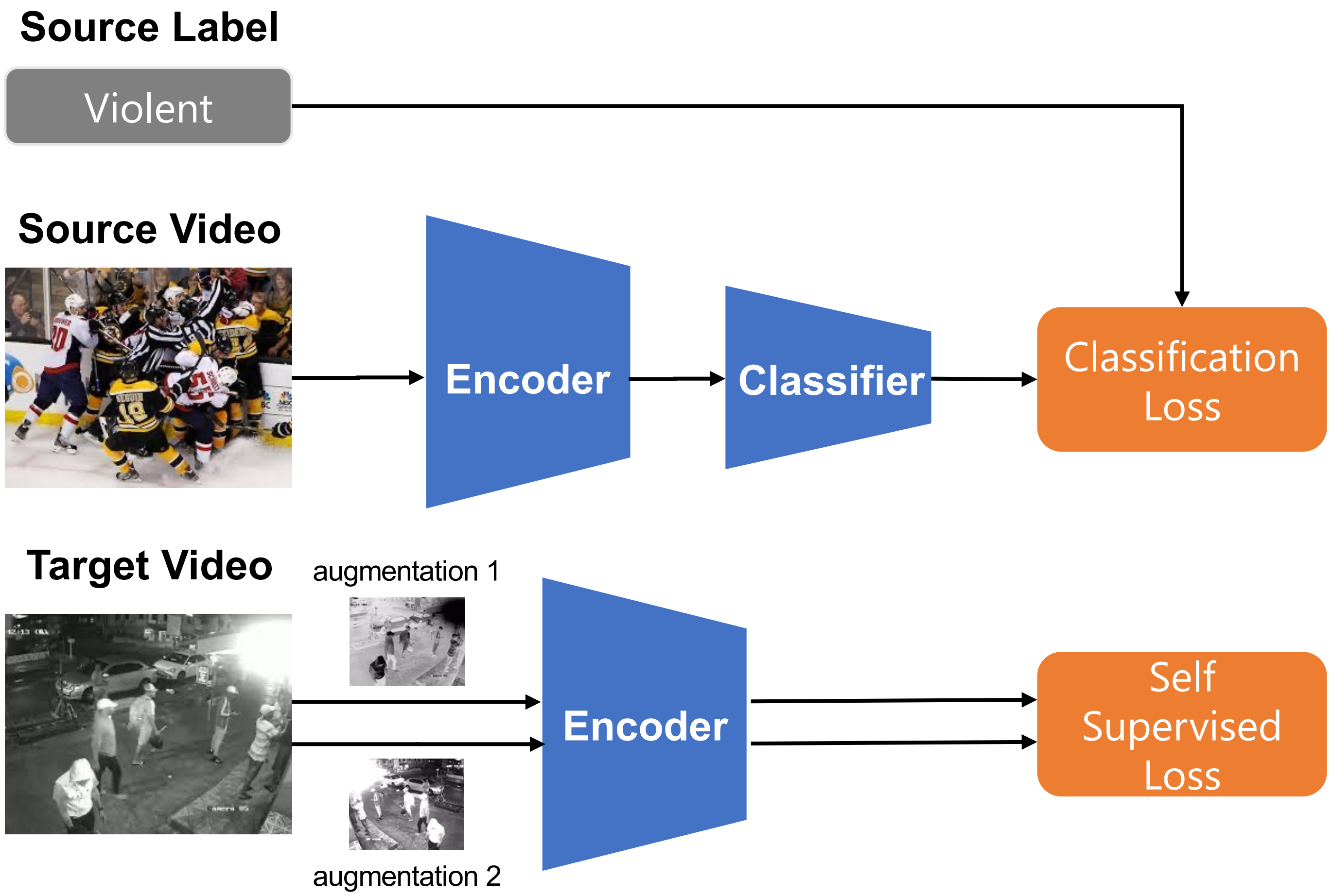
- Effective task performance in source domain due to labeled training data
- Self-Supervised Learning is used to learn a good feature space for the target domain

→ Good Representation space for **both task and target**

**Choosing a self-supervised task:**

1. **Contrastive Method**: Good performance but need **large batches and negative pairs**
2. **Non-Contrastive Method:** No need for negative pairs and large batches but **representation collapse risk**
3. **Maximum-Entropy Coding**: **Combination** of both contrastive and non-contrastive leading to robust performance with **small batch size**

# Self-Supervised Learning To Learn Good Representations for Target Domain




**The Approach:**

- Effective task performance in source domain due to labeled training data
- Self-Supervised Learning is used to learn a good feature space for the target domain

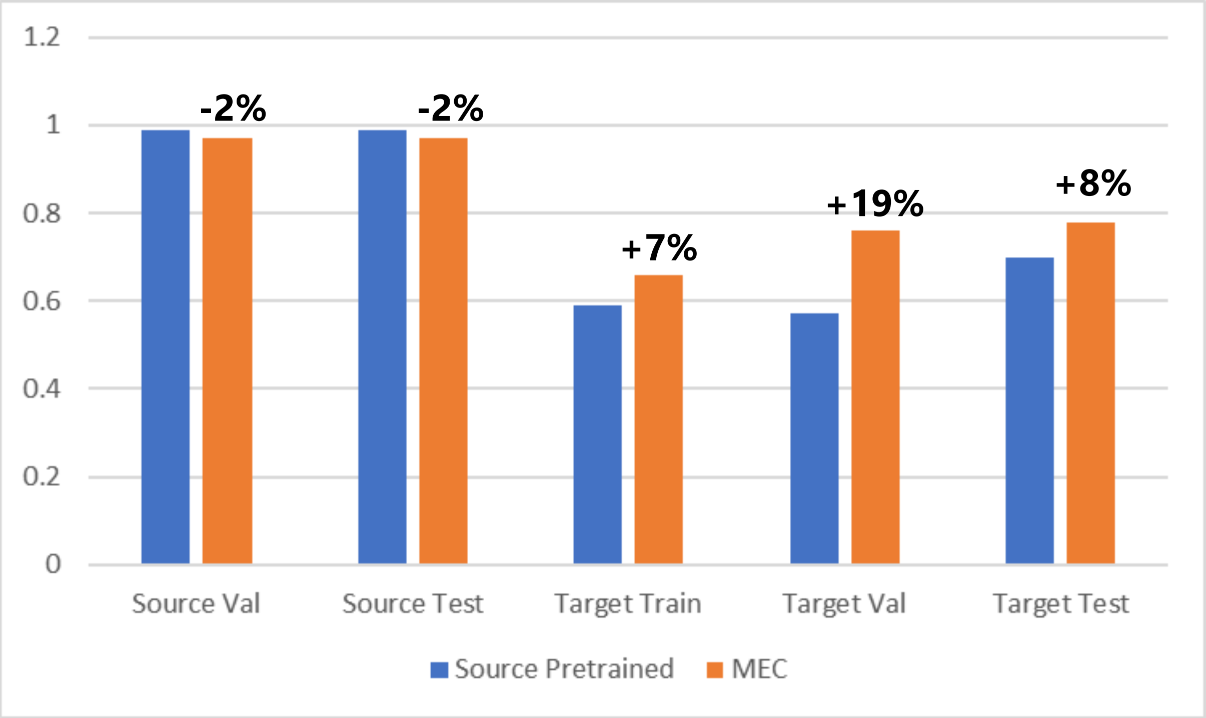
→ Good Representation space for **both task and target**

**Choosing a self-supervised task:**

1. **Contrastive Method**: Good performance but need **large batches and negative pairs**
2. **Non-Contrastive Method:** No need for negative pairs and large batches but **representation collapse risk**
3. **Maximum-Entropy Coding**: **Combination** of both contrastive and non-contrastive leading to robust performance with **small batch size**

# Self-Supervised Learning Improves Target Domain Performance

**Performance Values Train, Validation, and Test Set**
[All values measured in % accuracy]




- SSL(MEC) **improves performance in the target domain**
- **Source Domain Performance degrades slightly**
- **Ablation:** Better than dividing into target encoder pretraining and source domain classifier fine-tuning into separate stages. Hypothesis:
  - The encoder is not suitable for the source domain due to the domain gap when pretrained separately
  - Features more related to violence are likely to be learned more.

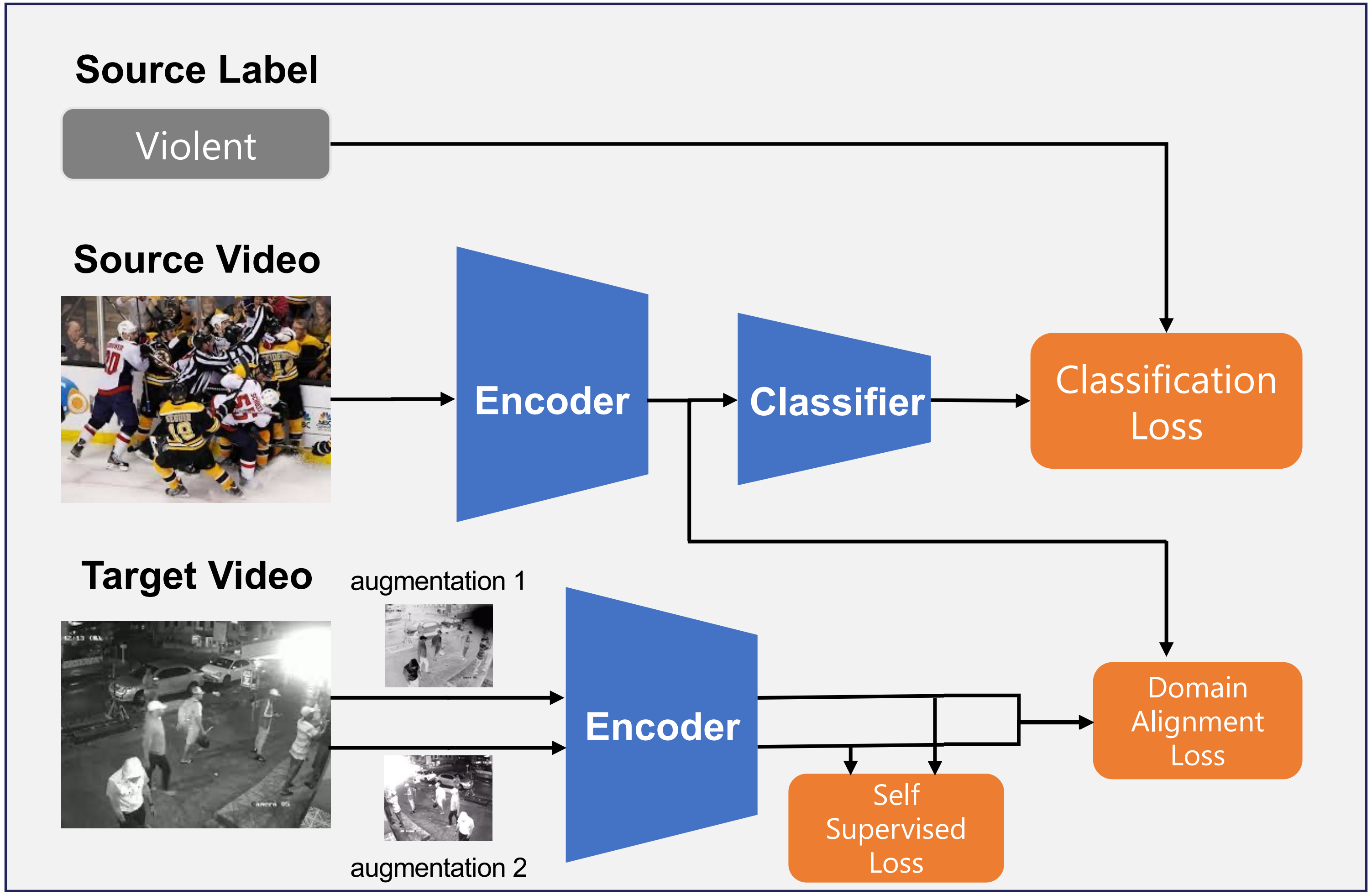
# UDA and SSL are Complementary

| | |
|---|---|
| **Method 1: UDA** | **Strength:** Alignment between source and target domain<br><br>**Weakness:** Difficulty in learning diverse features from the target domain |
| + | |
| **Method 2: SSL** | **Strength:** Ability to learn diverse features from the target domain<br><br>**Weakness:** Lack of alignment between source and target domains |

**UDA and SSL are Complementary**

- UDA & SSL have separate strengths & weaknesses
- Let's mix the two methods into one to see if they each cover up other's weakness.

# UDA + SSL to learn the best representation space for task on target domain




**The Approach:**

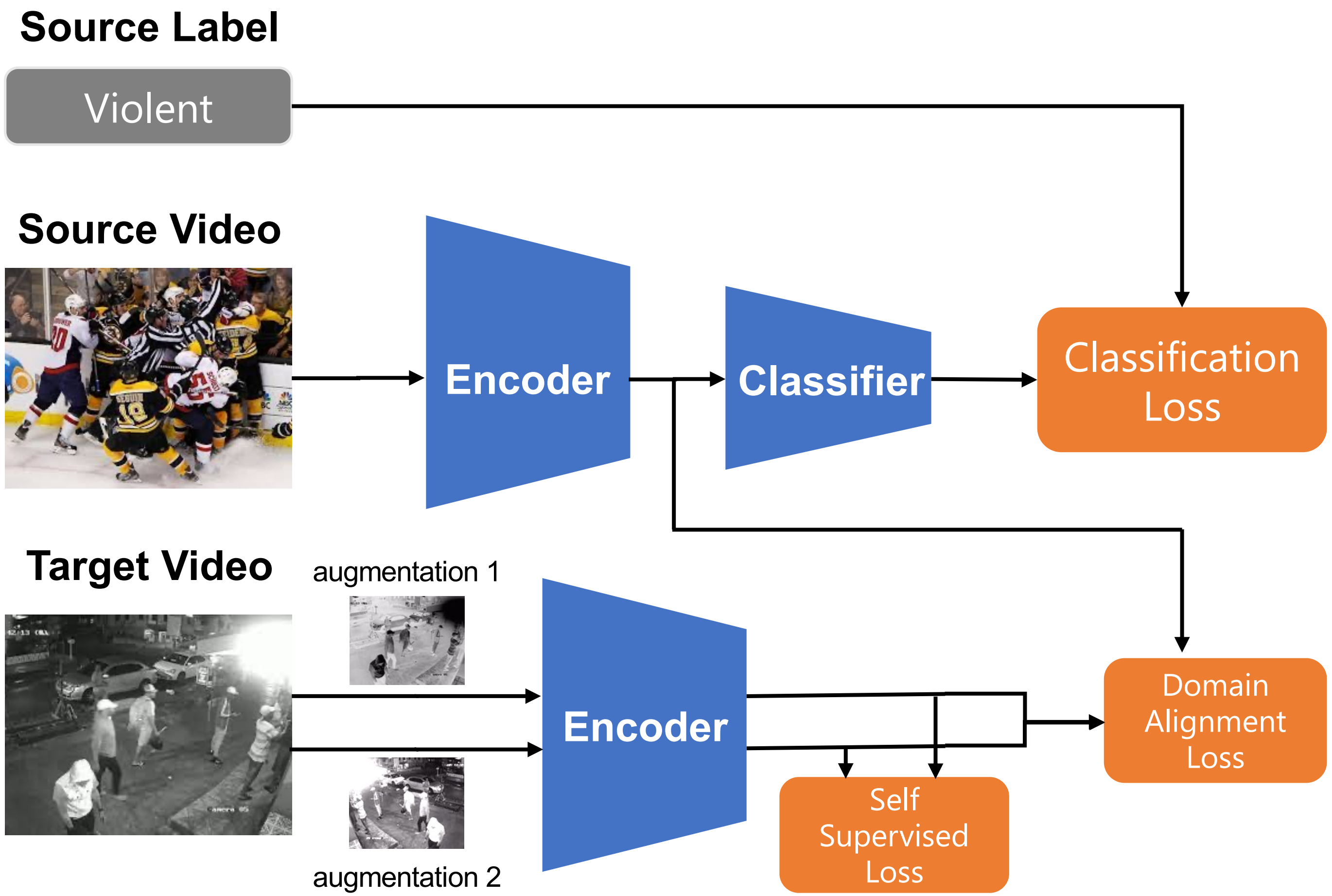
- Align domain between source and target
- Learn rich representations for target domain

→ Achieve the best representation space for task, source, and target

**Choose Best Performing methods:**

1. **Wasserstein Loss:** Better performance over Maximum Mean Discrepancy Loss
2. **Maximum-Entropy Coding**: Works well with small batch sizes

# UDA + SSL to learn the best representation space for task on target domain




**The Approach:**

- Align domain between source and target
- Learn rich representations for target domain

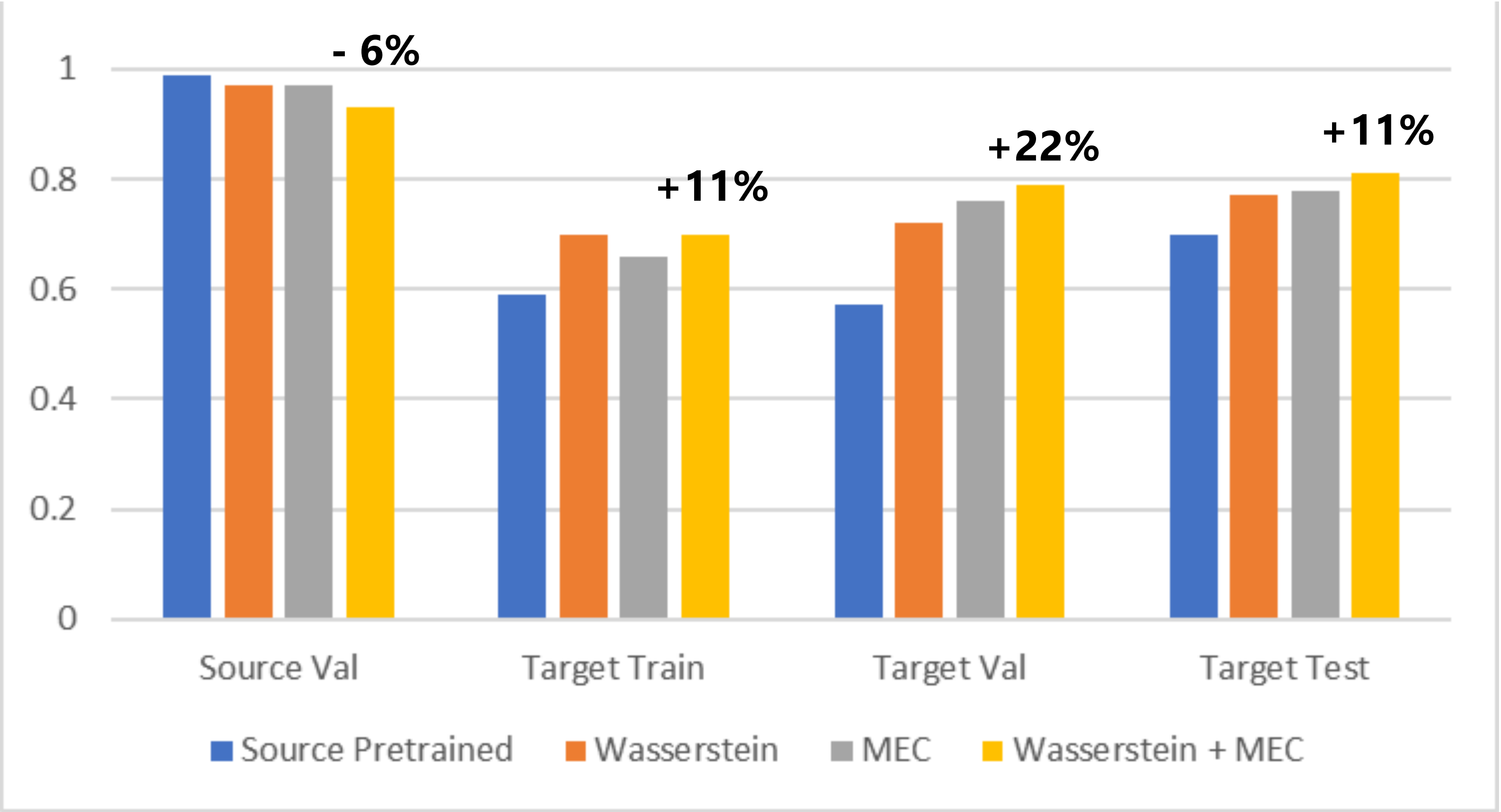
→ Achieve the best representation space for task, source, and target

**Choose Best Performing methods:**

1. **Wasserstein Loss:** Better performance over Maximum Mean Discrepancy Loss
2. **Maximum-Entropy Coding**: Works well with small batch sizes

# UDA + SSL Improves Target Domain Performance

**Performance Values Train, Validation, and Test Set**
[All values measured in % accuracy]




- Mixing UDA and SSL **improves over the methods applied individually**
- Source domain performance **degrades by over 6%**, indicating that **source domain overfitting** has decreased
- The source-target **domain accuracy gap** decreased from **29% to 12%**

D

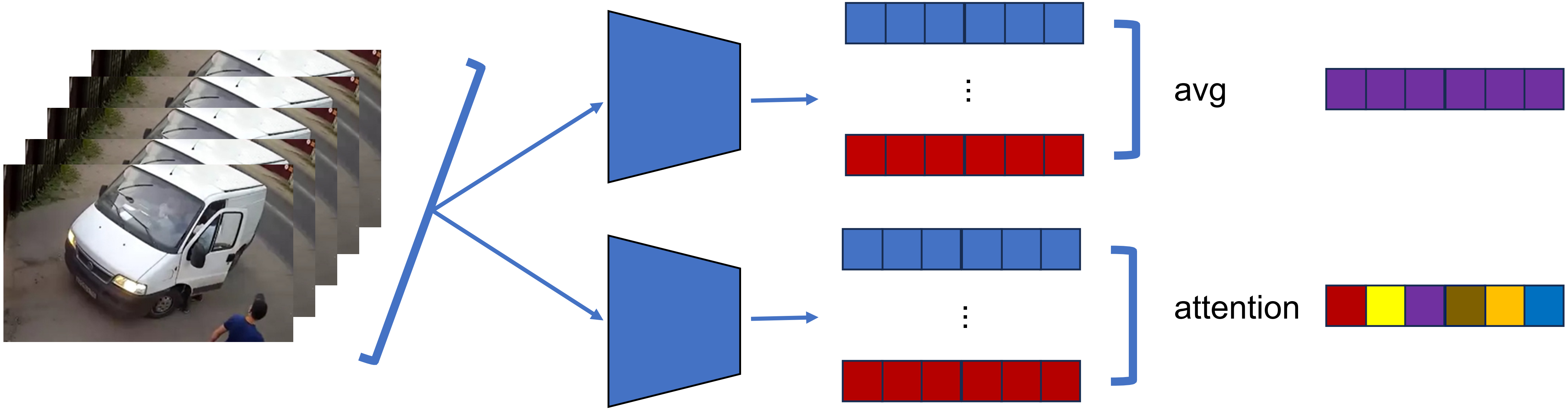
# Future Work

# Can Temporal Attention Help to Improve Latent Space Representation With More Information?

- Unsupervised Domain Adaptation
- Self-Supervised Learning

⇒ Both done in the **latent representation space**. (align, learn rich features for the representation space)

But video has time axis. Can't **temporal information** be more used for UDA or SSL? ⇒ **Attention**?



E

# Conclusion

# Conclusion

- Targeted **realistic unsupervised domain adaptation** challenge for CCTV footage.
- **Adapted unsupervised domain adaptation** and **self-supervised learning** methods normally used in the image domain to the **video domain**.
- Analyzed design choices, the reason behind it, leading to an **8% improvement**.
- **Combined methodologies** (UDA + SSL) boosted performance by an **additional 3%.**
- **Future Direction:** Exploring **temporal attention** mechanisms for further enhancements.

F

# Appendix

# MoViNet

**The Challenge**

- CCTV cameras operate on non-high-performance devices
- Violence detection in real time

**MoViNet**

- Very efficient general-purpose network for video analysis
- Can easily be run on **“The Edge”** or on **smartphones**
- Good baseline performance